Technical Manual for the Teaching Strategies *GOLD*® Assessment: Birth Through Third Grade

Richard G. Lambert, Ph.D.

Center for Educational Measurement and Evaluation UNC Charlotte

June, 2026

3rd Edition

TeachingStrategies®

Table of Contents

Overview of *GOLD*®

GOLD® is a developmental child assessment system intended to provide formative and summative data on the whole child from birth through third grade. It enables teachers to collect documentation on an ongoing basis to measure children's knowledge, skills, and abilities along a series of developmental progressions. In contrast to direct assessments—which are more constrained in what and how they measure—*GOLD*® is designed for teachers to regularly collect evidence of child development during authentic interactions and activities that take place in natural settings, including classrooms, playgrounds, homes, and more. In addition, the process gives young children agency by authentically involving them in meaningful interactions that support developmental progress.

With *GOLD*®, assessment is an ongoing process of compiling portfolios of evidence for each child, reflecting upon and analyzing the evidence, assigning preliminary levels, and finalizing "checkpoint" ratings at regular intervals during the year. The primary intent of the assessment is to help teachers observe and understand child progress, plan instruction, and support child growth and development. Teachers and administrators may also use *GOLD*® data to individualize children's early education and care in co-collaboration with caregivers, case managers, and clinicians; guide programming for children and families; and inform professional development content for staff.

Purpose of the Technical Manual

It is vital to children's growth and development to provide assessments that yield reliable, valid, and culturally sensitive information. Reliability and validity are not inherent qualities of an assessment itself but rather are properties of the information an assessment provides under particular conditions of use. Hence, the measurement properties of any assessment should be rigorously examined—for all subgroups of children and conditions of use.

Aims of the Revisions

This is a revised edition of the second technical manual for the birth through third grade (B–3) version of *GOLD*®; the first B–3 manual was released in 2017 alongside the expansion into grades 1–3; the second B–3 manual was released in 2020 to provide updated norms. It is important to note that nothing about the *GOLD*® assessment has changed since the release of the second manual in 2020. Instead, the aims of the revisions are as follows.

1. Enhance the accessibility of information about *GOLD*® by summarizing, in one publicly accessible document, existing evidence for the validity and reliability of child development data collected with the tool
2. Update the field about new evidence supporting the validity and reliability of *GOLD*® data for specific subgroups of children
3. Provide new post-pandemic norms based on data from the 2023-2024 school year

The Design of *GOLD*®

Taking a whole-child approach, *GOLD*® assesses children's development and learning across four **developmental domains** (social–emotional, physical, language, cognitive) and five **content domains** (literacy, mathematics, science and technology, social studies, and the arts). It also includes a tenth domain, English language acquisition, for use with multilingual learners (MLLs). Each domain comprises a set of **objectives** designed to guide teachers through the assessment process. Many objectives are further broken down into two or more **dimensions**. *GOLD*® has 38 objectives in total, collectively termed the "objectives for development and learning" (ODL). Figure 1 illustrates the organization of the ODL for the physical domain as an example.

In addition, each dimension under the first six domains (social–emotional, physical, language, cognitive, literacy, and mathematics) has a progression associated with it. Every **progression** represents a continuum that enables teachers to relate observable behavior to expectations for a child's age/grade. Progressions help teachers and other stakeholders understand how children are growing relative to developmentally appropriate expectations. Figure 2 illustrates an example of the progression for Objective 4 under the physical domain.

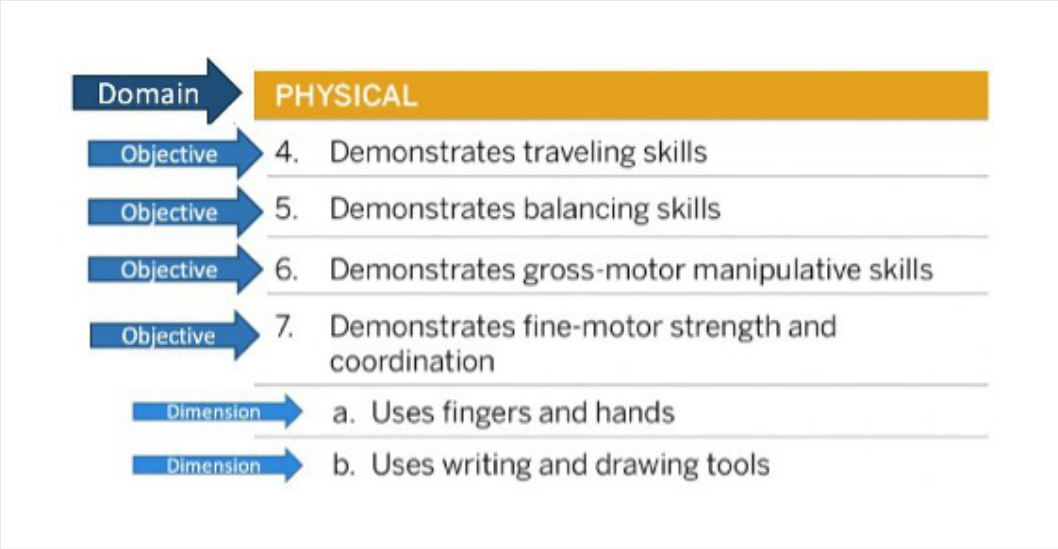

Figure 1. ODL for the Physical Domain

Objective 4: Demontrates traveling skills

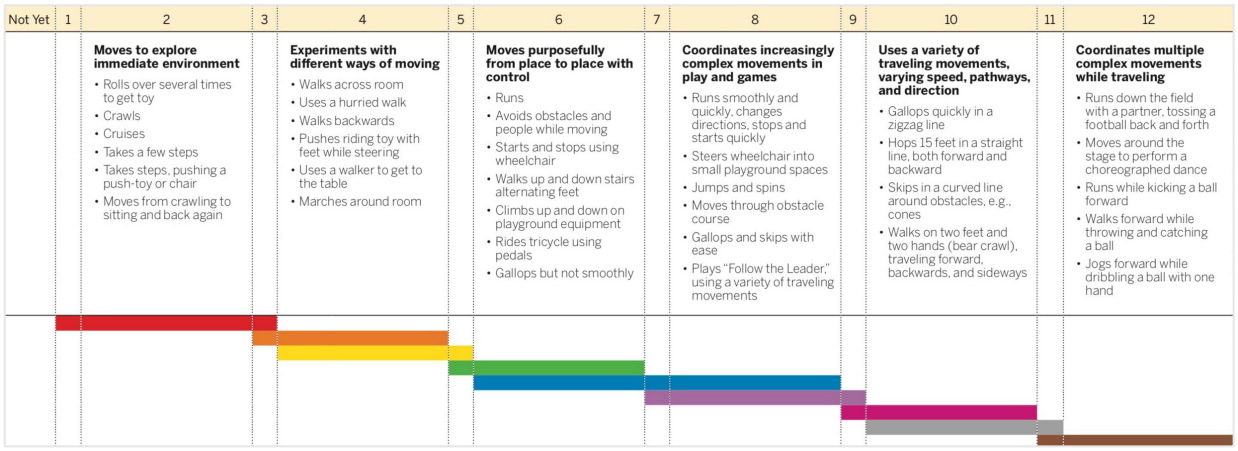

Figure 2. Progression Example: Objective 4

As exemplified above, *GOLD*® uses color bands to represent the associated developmental and learning expectations for a given year of life or academic program year. Each band corresponds to the range of **widely held expectations (WHE)** for a particular age/grade. Knowledge, skills, and abilities that fall within a child's respective color band are meeting widely held expectations for that objective or dimension. Those that fall to the left of the color band are below expectations, and those that fall to the right are exceeding expectations. Color bands that fall completely under the "Not Yet" level indicate that it is not yet developmentally appropriate to expect children to demonstrate the skill(s). WHE were developed by panels of experts in child development based on the latest developmental theory and research. This manual focuses on the six domains that have progressions for which WHE have been developed.

Figure 3. GOLD® Color Bands

The Purpose of *GOLD*®

Users must understand the intent of any assessment to ensure that they actually assess the intended outcomes. Therefore, it is valuable to delineate appropriate and inappropriate uses. *GOLD*® has been designed and externally validated for use as a formative, developmental, authentic, and criterion-referenced assessment. Its primary purpose is to provide teachers with instructionally relevant information about the children they teach. Figure 4 contrasts optimal and ineffective uses of *GOLD*®, and the sections below it describe the most meaningful applications of the information it generates.

Optimal Uses of *GOLD*®

GOLD® Is a Formative Assessment	• Formative assessment focuses on the learning process and is used to support learning while learning is taking place. • Formally, formative assessment is "a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to help students improve their achievement of intended instructional outcomes" (CCSS, 2006; AERA/APA/NCME, 2014).
GOLD® Is a Developmental Assessment	• *GOLD*® includes progressions of growth, development, and learning that describe a sequence of stages and behavioral anchors that children are generally expected to demonstrate at a given age/grade. • It is designed to help teachers assess the whole child.
GOLD® Is an Authentic Assessment	• Authentic assessments help teachers observe the progress children make through a process that emerges naturally. • Evidence of development and learning is gathered during everyday instructional activities and routines.
GOLD® Is a Criterion-Referenced Assessment	• *GOLD*® measures developmental progress and learning relative to a fixed set of standards, i.e., the ODL. • One child's performance on *GOLD*® does not depend on all other children's.

Figure 4. Optimal and Ineffective Uses of GOLD®

GOLD® Is a Formative Assessment

GOLD® is a valuable resource for teachers getting to know children at the beginning of the school year. It can help them understand the strengths that a child brings to the classroom as well as areas where a child needs support. It is also useful at any point in the academic year, allowing teachers to understand the current status of the growth, learning, and development of the children in their classrooms. Finally, it can help teachers identify children's interests, plan classroom activities, select and rotate classroom materials, and individualize and differentiate instruction.

When teachers communicate with families and other guardians, *GOLD*® can help them do so in terms that can be easily accessed and understood. Teachers can link specific examples of what a child knows and can do with placements on the developmental progressions. This process can help facilitate rich conversations about the child's development within the classroom, family, and cultural context. Teachers can even solicit evidence of child progress and development from families and other caregivers to inform the assessment process and ensure that progress is measured accurately. This process can help teachers partner with families to support the growth and development of each child.

Other advantages of *GOLD*® include helping teachers collaborate with other educational professionals who serve the same children, identifying areas of need for professional development and other training, empowering mentors to better support teachers with planning instructional activities, and enhancing observational skills and awareness of how children learn and function in the classroom. In summary, *GOLD*® is especially effective at supporting data-driven decisions around individualizing support for children.

GOLD® Is a Developmental Assessment

GOLD® has been aligned to the Head Start Early Learning Outcomes Framework, the Common Core Standards, and state early learning standards. Most standards outline specific skills and abilities that children are expected to obtain by the end of a particular grade or age level.

GOLD® expresses these standards not just in terms of the end point, but also in terms of the journey or learning process. For each standard, a set of instructional objectives has been outlined, and, in turn, each instructional objective is associated with at least one developmental progression. Each progression details the steps children are expected to follow on their way to mastery of new skills and abilities. In this way, the assessment helps teachers connect and understand the relationship between learning standards, curricular goals, and instructional objectives in terms of evidence that can be observed within everyday classroom contexts. Most importantly, it helps teachers individualize the implementation of scaffolds each child needs to reach the next step of their developmental trajectory.

GOLD® Is an Authentic Assessment

Authentic assessments help teachers observe the progress children are making through a process of gathering evidence of learning as it emerges naturally throughout the course of typical classroom activities. Evidence documented by teachers is then used to assess where children are on a progression and support further development along the progression. In this way, *GOLD*® supports assessment "for" learning and assessment "about" the learning process, not just assessment "of" the results of learning (Heritage, 2013). The authentic process of formative assessment has often been described as a continuous cycle of activities that are part of everyday instruction (Heritage, 2013). This cycle is often outlined in phases, as illustrated in Figure 5.

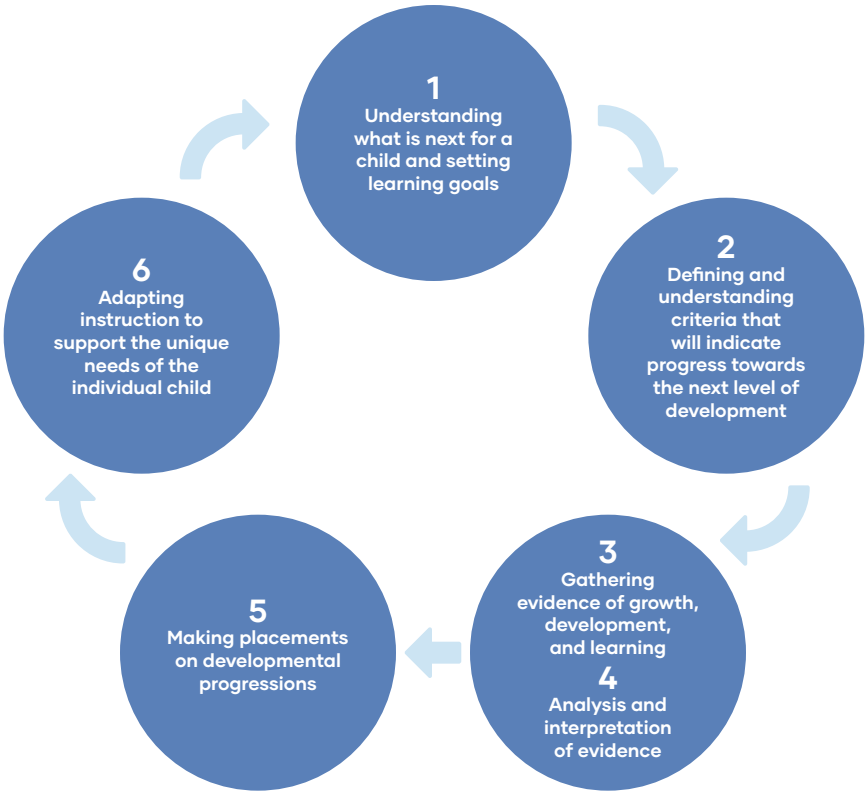

Figure 5. The Authentic Assessment Cycle

This cycle repeats itself as the child moves toward the next developmental level on a specific progression related to an instructional objective. In the end, teachers and children alike benefit from an assessment process that not only avoids disrupting regular classroom activities but also supports development and learning as assessment takes place.

GOLD® Is a Criterion-Referenced Assessment

The information provided by *GOLD*® is most useful for identifying where a child is developmentally relative to their own past developmental trajectory and to standards for children of a given age range. Teachers can use the widely held expectations for each color band to understand which behaviors and skills children can generally be expected to demonstrate in the classroom. These expectations are not grounded in quantitative norms that describe how children of a given age have scored on the measure at a fixed point in time.

Rather, they are based on developmental theory, expert recommendations, and empirical research. While criterion-referenced assessments are not norm-referenced assessments, the norms that are available toward the end of this manual are designed to provide teachers with an additional resource. Ultimately, the norms are for those who are interested in a broad and comprehensive picture of how a child is growing and developing relative to the developmental progress of other children across the nation.

Using the Summative Scores Provided by the *GOLD*® Assessment System

Despite its primary purpose as a formative assessment, the *GOLD*® assessment system also provides supplementary summative scores. There are several types of scores available for the six core *GOLD*® domains: social–emotional, physical, language, cognitive, literacy, and mathematics. These include **raw scores**, **widely held expectations scores**, **scaled scores**, and **national norm scores**. Each of these scores will be described in this section of the manual, along with guidelines for their appropriate use.

Raw Scores

GOLD® raw scores are calculated from the sum of all the finalized ratings a teacher made on the progressions under each domain. Raw scores, hence, represent a child's current knowledge, skills, and abilities related to a particular domain of development. At the end of each checkpoint, a child may have up to six different raw scores.

Widely Held Expectations Scores

Widely held expectations reflect the expected developmental trajectories for children on each dimension. Each color band corresponds to the knowledge, skills, and abilities children are expected to demonstrate over a given year of life or from the beginning to the end of a program year. As already mentioned, if a child's rating falls within the color band that corresponds to her age/grade, that child's skills are meeting WHE for the dimension. If her rating falls to the left of the color band, her skills are below expectations; if it falls to the right, her skills are exceeding expectations. Generating *GOLD*® reports that display WHE in *SmartTeach*™ enables teachers to compare data for specific children or groups of children to determine whether their knowledge, skills, and abilities are below, meeting, or exceeding expectations. *SmartTeach*™ offers teachers the ability to make these comparisons at the level of individual dimensions and at the domain level.

To find the range of WHE for a given domain, sum the numeric values on each progression that correspond to the bottom level of each color band; this total assigns the lower bound of the domain's WHE. Then, sum the numeric values on each progression that correspond to the uppermost level of each color band. Similarly, this total assigns the upper bound of the domain's WHE. For example, for Pre-K 4 (the blue color band), WHE for the physical dimensions range from 6–8 for Objective 4, from 6–8 for Objective 5, from 6–8 for Objective 6, from 6–8 for dimension Objective 7a, and from 5–7 for Objective 7b. Therefore, the range of widely held expectations for the blue band is (6 + 6 + 6 + 6 + 5) to (8 + 8 + 8 + 8 + 7), or 29–39. If the raw score for a Pre-K 4 child's physical abilities is 33, which falls within the range of 29–39, her performance meets WHE for the physical domain. Tables 16 through 21 present the proportion of children who are below, meeting, and exceeding WHE in the fall, winter, and spring based on a nationally representative sample for each age/grade.

Scaled Scores

It is important to recognize that raw scores are intended for use only in generating domain-level WHE scores. Raw scores alone are not useful for tracking growth and development over time for the following reasons. First, the progressions within a given domain of development do not include the same number of steps and, therefore, do not contribute the same number of potential points to the total raw score. Second, the raw scores for each progression represent an ordinal level of measurement. This means that a score of 2 represents a higher level of development than a score of 1 and a lower level of development than a score of 3, etc.

However, these scores cannot be interpreted such that an increase of one point represents an equal amount or "quantity" of development. In order to facilitate the interpretation of a child's growth over time in terms of equal intervals, or score points, raw scores must be converted to scaled scores.

GOLD® has been designed and validated to be used primarily as a criterion-referenced, authentic, formative assessment. Nevertheless, two types of scores are provided to users for the purpose of specific norm-referenced interpretations: scaled scores and national norm scores. Developmental scaled scores are provided for the purposes of converting raw scores from each domain of development to a common scale. These scaled scores eliminate the impression that one domain of development is more important than another. Scaled scores are primarily provided for the purpose of tracking a child's growth and development across different age/grade levels, i.e., they allow teachers to compare the amount of growth a child made over different time periods. For example, scaled scores can tell a teacher whether a child grew more/less over the spring checkpoint period than she did over the fall checkpoint period or whether she grew more/less throughout one school year versus another. Scaled scores also allow for comparisons between children, even if they are at different levels of development. Note that comparisons may only be made between scaled scores from the same domain: **scores from different domains are not comparable**. The *GOLD*® uniform scale ranges from 0 to 1,000.

National Norm Scores

GOLD® scaled scores are also used to create national norm scores to facilitate interpretations of how children of a particular age range tend to be rated by other teachers across the nation. National norm scores are based on quartiles, or fourths, of the national distribution of scaled scores for each color band for a given domain. Children with scaled scores in the first quartile (lowest 25%) are considered below the national norm, children with scores in the second and third quartiles (the middle 50%) are considered within the national norm, and children with scores in the fourth quartile (highest 25%) are considered above the national norm. These scores can be a valuable resource for teachers interested in understanding how the children they serve have been placed on the progressions relative to how teachers across the nation place the children in their own classes. To that end, updated national norms are provided in Tables 10 through 15.

Guidelines for Aggregated Reporting

The most important considerations with regard to reporting assessment scores are 1) the purpose, 2) the specific inferences that are to be made from the scores, and 3) whether these align with the nature of the assessment in question. With respect to *GOLD*®, the assessment was designed—and has been extensively validated—as an authentic, developmental, formative assessment. Its primary purpose is to provide actionable information for teachers as they plan instruction and support the growth and development of young children.

It is possible to use aggregated *GOLD*® data in a responsible way. However, if users are considering aggregated reporting, various threats to the validity of the inferences made from the scores need to be given careful consideration. The teachers producing such data must have received all appropriate training, including ongoing coaching. It is also preferred that they have achieved inter-rater reliability certification. Reliable assessors are just as likely to be strict as they are to be lenient in their placements of children on the developmental progressions. Still, some teachers' placements may be consistently strict or lenient. Nevertheless, with adequate training, any measurement error caused by rater effects should wash out in the aggregate. It is essential that this assumption holds to ensure responsible use of aggregated scores on *GOLD*®.

About Missing Data

GOLD® requires teachers, as part of the ongoing assessment process, to collect evidence of each child's developmental progress during the normal course of classroom activities. For some assessment periods, teachers may not have the opportunity to collect sufficient evidence for a child to support a finalized placement on a particular progression. Some children join a classroom later in the assessment period; others miss classroom activities that generate evidence pertaining to a given progression. Since placements (or "ratings") tend to be highly correlated with each other within a given

domain of development, imputation of reasonable amounts of missing data can allow a total score to be obtained. Note that any imputation is domain-specific, i.e., when a progression is missing a rating, only progressions from within the same domain are used to impute its value.

Teaching Strategies has developed an imputation method that takes into account not only the varying number of steps on the progressions but also the varying locations and widths of the color bands. Termed *tripartite weighting*, this method involves using each rating's relative placement with respect to the total number of steps in its corresponding region of the progression to impute missing ratings. Progressions are divided into three regions relative to each color band: below the band, within the band, and above the band. Each region may have a different number of steps, depending on the developmental level of the progression and the number of steps in the overall progression. If a child receives a rating of 3 that happens to fall in the region below a color band that begins at 6, for example, then the rating's relative placement on the region below the color band is 0.6. Likewise, if a child receives a 5 and the rating falls within a color band that spans scores 4 through 8, then the rating's relative placement on the region within the color band is 0.25. Any missing rating is estimated once per relative placement of all the ratings that are present. The average of these estimations yields the imputed rating for a given progression. For instance, if a child is missing 1 out of 10 ratings for a given domain, then that progression's imputed rating will result from the average of 9 estimations; if a child is missing 2 out of 10 ratings, then both progressions' imputed ratings will result from separate averages of 8 estimations. Total raw scores may be calculated by summing the original and imputed ratings and rounding down the result. *GOLD*® scaled scores may then be calculated as normal from the resulting raw scores.

National Sample

The sample for norming was drawn from a population of 542,694 children who were assessed on every *GOLD*® dimension during fall, winter, and spring of 2023–2024. Due to the limited sample size in kindergarten through third grade, an additional 5,416 children from the 2022-2023 school year were added to the pool for these grade levels. Teachers in programs that assess children in all three semesters are more likely to understand the process of collecting valid evidence as well as how to translate that evidence into meaningful placements on the developmental progressions. In other words, their ratings are likely more reliable than ratings assigned by teachers who assess children only once or twice per year. Nevertheless, a few qualifiers about the sample should be mentioned. Not all of the programs have adopted the *GOLD*® inter-rater reliability certification process, and others are in various stages of requiring and/or achieving this certification for all of their teachers. In addition, teachers working with children who are in kindergarten through third grade used the curriculum that is outlined for them by their local education agency. Therefore, they do not use *The Creative Curriculum*®. Conversely, many of the programs serving younger children were users of *The Creative Curriculum*®.

A stratified random sample of 5,000 children was drawn from each birth to kindergarten age/grade group (color band). The strata were formed using the U.S. Census Bureau's data for race/ethnicity in an effort to represent each subgroup in proportion with the U.S. population of children. Incidentally, this process mitigates any clustering of children within their rater or teacher, as whole classrooms of children were not selected. The *GOLD*® assessment has not been adopted for use in first, second, and third grades at the same rate as it has for the birth to kindergarten years. Therefore, all available data from children in first, second, and third grade who have fall, winter, and spring assessment records was used, regardless of school year. This process, as shown in Table 1, resulted in a total norm sample of 31,751 children.

Given that Hispanic identity is an ethnicity, not a racial grouping, and given the importance of representing children of Hispanic ethnicity in the norm sample, the race and ethnicity variables were combined into the following six race/ethnic subgroups: 1) White, not Hispanic; 2) African-American, not Hispanic; 3) Native American or Hawaiian/Pacific Islander, not Hispanic; 4) Asian, not Hispanic;

5) multiracial/other, not Hispanic; and 6) Hispanic. As shown in Table 2, the norm sample was roughly balanced by gender (boys = 50.29%, girls = 49.7%). Children with an IFSP or IEP comprised 7.63% of the sample. The primary language spoken in the home was distributed as follows: English (79.88%), Spanish (12.66%), and other languages (7.46%). The race/ethnicity of the children in the sample was as follows: a) White – 46.23%, b) African American – 15.04%, c) Native American/Pacific Islander – 1.92%, d) Asian – 4.16%, e) Multiracial/other – 5.04%, and f) Hispanic – 26.9%. These values indicate that the sample was approximately nationally representative for all race/ethnicity groups of American children. A total of 22.19% of the norm sample qualified for the National School Lunch Program (free or reduced-price lunch) as reported by their teacher. This figure is likely an underestimate as teachers may not accurately report this information. As is, this number reflects an under-representation of these children, as the national figure is over 50%.

Furthermore, the sample of children from birth to kindergarten received educational services in programs that were located in all 50 U.S. states and the District of Columbia, Puerto Rico, and limited locations around the world that serve the families of U.S. military personnel. The sample includes a rich geographic diversity from all regions of the U.S. and includes all types of settings where young children are served (Head Start, private prekindergarten, publicly funded preschool programs, and public schools).

Item Response Theory Analyses Related to the Construction of Scaled Scores

The measurement properties of *GOLD*® were analyzed using assessment records from the 2018–2019 academic year. Over 1.5 million American children were assessed at some point during the year, providing a broad pool from which to draw a nationally representative sample. Study methods involve the use of Rasch measurement models to examine the properties of the information provided by each developmental progression. Specifically, results are first presented for dimensionality to confirm the appropriateness of the model, followed by analysis of rating scale validity, item difficulty measures, and reliability measures. Note that for the remainder of this manual we will use the terms *item* and *progression* interchangeably.

Rasch IRT models were used to create ability estimates for each child on each domain and to examine the measurement properties of the information provided by each developmental progression. *GOLD*® assessment data were analyzed using the Rasch partial credit model (PCM) (Masters, 1982) with Winsteps software (Linacre, 2012). A separate Rasch analysis was conducted for each of the six domains of development. The rating scale model (RSM; Bond & Fox, 2001) and the PCM are the two most widely used Rasch models for polytomous response data. The PCM, rather than the RSM, was chosen because the progressions do not share the same rating scales (i.e., the number of steps is different across progressions). Specifically, 11 *GOLD*® progressions are on a 0–9 scale, 6 are on a 0–11 scale, 17 are on a 0–13 scale, 25 are on a 0–15 scale, and 1 is on a 0–19 scale. For each progression, the 0 category represents "Not Yet," and the maximum category represents abilities beyond the most developmentally advanced behavioral anchor.

Dimensionality

Rasch modeling assumes unidimensionality, meaning that the items under a given domain measure only one latent construct. The unidimensionality of each scale was evaluated via mean-square (MNSQ) fit statistics and principal components analysis of Rasch residuals (PCAR). MNSQ fit values between 0.6 and 1.4 are considered reasonable for rating scale items (Bond & Fox, 2007). MNSQ values less than 2.0 can indicate that an item, though not fitting optimally with the measurement model, can still contribute useful information to the overall score on the measure. In other interpretations, items with mean square values between 1.4 and 2.0 can be considered potentially unproductive for the development and construction of new measurement scales but not degrading to the quality of the information provided by the scale (Linacre, 2002). Infit statistics

indicate the fit of individual item response patterns to the measurement model. They also address fit to the underlying construct and the possibility of secondary dimensions. Outfit statistics are sensitive to outliers, that is, responses that show great differences between person responses and item difficulties. They are also sensitive to unusual and unexpected item response patterns. For PCAR, a variance of greater than 50% explained by measures is considered good and offers support for scale unidimensionality. If a secondary dimension has an eigenvalue of less than 3 and accounts for less than approximately 5% of the unexplained variance, unidimensionality is considered plausible (Linacre, 2012). The following sections present the results of the MNSQ fit and PCAR analyses; note that indices were evaluated for the fall assessment checkpoint.

Social–Emotional Scale (9 progressions)

The PCAR showed that for the social–emotional scale, the Rasch dimension explained the majority of the variance in the data (91.9%) with an eigenvalue of 101.4. The first contrast (the largest secondary dimension) had an eigenvalue of 1.9 and accounted for 1.8% of the unexplained variance. All fit statistics for all of the social–emotional progressions were within acceptable limits. The infit MNSQ values ranged from 0.85 to 1.18. The outfit MNSQ values ranged from 0.83 to 1.20. The item total score correlations ranged from .93 to .95.

Physical Scale (5 progressions)

The PCAR showed that for the physical scale, the Rasch dimension explained the majority of the variance in the data (92.9%) with an eigenvalue of 65.0. The first contrast had an eigenvalue of 1.6 and accounted for 2.2% unexplained variance. All fit statistics for all of the physical progressions were within acceptable limits with one exception. The infit MNSQ values ranged from 0.84 to 1.22. The outfit MNSQ values ranged from 0.81 to 2.22. The progression for Objective 7b had an outfit statistic just outside the optimal range (2.22), suggesting the presence of outliers or unusual response patterns. The item total score correlations ranged from .94 to .96.

Language Scale (8 progressions)

The PCAR showed that for the language scale, the Rasch dimension explained the majority of the variance in the data (93.4%) with an eigenvalue of 113.8. The first contrast had an eigenvalue of 1.5 and accounted for 1.2% of the unexplained variance. All fit statistics for all of the language progressions were within acceptable limits. The infit MNSQ values ranged from 0.78 to 1.15. The outfit MNSQ values ranged from 0.79 to 1.17. The item total score correlations, excluded from the total score, ranged from .94 to .96.

Cognitive Scale (10 progressions)

The PCAR showed that for the cognitive scale, the Rasch dimension explained the majority of the variance in the data (93.3%) with an eigenvalue of 140.4. The first contrast had an eigenvalue of 1.9 and accounted for less than 1.2% of the unexplained variance. All fit statistics for all of the cognitive progressions were within acceptable limits. The infit MNSQ values ranged from 0.79 to 1.07. The outfit MNSQ values ranged from 0.79 to 1.12. The item total score correlations ranged from .93 to .96.

Literacy Scale (16 progressions)

The PCAR showed that for the literacy scale, the Rasch dimension explained the majority of the variance in the data (92.2%) with an eigenvalue of 189.8. The first contrast had an eigenvalue of 2.0 and accounted for 1.0% of the unexplained variance. All fit statistics for all but two of the literacy progressions were within acceptable limits. The infit MNSQ values ranged from 0.70 to 1.47. The outfit MNSQ values ranged from 0.68 to 4.78. The progressions for Objectives 18d (4.78) and 15c (2.03) had outfit MNSQ values just above 1.4, suggesting the presence of outliers or unusual response patterns. The item total score correlations ranged from .55 to .94.

Mathematics Scale (12 progressions)

The PCAR showed that, for the mathematics scale, the Rasch dimension explained the majority of the variance in the data (90.3%) with an eigenvalue of 112.0. The first contrast had an eigenvalue of 2.4 and accounted for 2.0% of the unexplained variance. All fit statistics for the mathematics progressions were largely within acceptable limits. The infit MNSQ values ranged from 0.78 to 1.37. The outfit MNSQ values ranged from 0.71 to 2.19. Two progressions had outfit MNSQ values above 1.4: Objectives 20d (2.19) and 22b (1.83). These values suggest the presence of outliers or unusual response patterns. The item total score correlations ranged from .57 to .94.

In summary, with the few exceptions noted above, the model fit statistics generally suggest that the data fit the Rasch PCM very well. The results indicate, namely, that the data satisfy the unidimensionality assumption of the Rasch model.

Rating Scale Category Effectiveness

The rating scale categories were also examined to provide insight into whether teachers use the instrument in the manner in which it was intended. Rating scale category effectiveness is one way to measure the validity of the developmental progressions. The averages of the ability estimates for each domain (based on the total raw scores) were examined for each group of children placed at a particular rating on a given developmental progression. The averages should advance monotonically with rating scale category values (Bond & Fox, 2007). Andrich thresholds (also called step calibrations) from the PCM were also analyzed (1978).

Thresholds parameterize, relative to item difficulty, the points at which a rater is equally likely to select one rating or its neighboring rating along the latent variable underlying the progression (Bond & Fox, 2007). Thresholds should also increase monotonically along the rating scale categories. This means that the probability that a child is placed on the next step on the developmental progression should always increase as the overall ability level of the child increases.

Social–Emotional Scale

Eight of the nine progressions include a scale that ranges from 0 to 13, and one ranges from 0 to 11. There were no disordered category thresholds. The observed sample averages for each response category generally advanced as expected across the rating scales with no disordered steps. For 7 of the 9 progressions there were no disordered average placements. However, for two of the progressions (Objectives 1a and 2a), there were disordered averages at the top of the scale. This means that the average total score for children placed at the top of the scale was below the average of those placed at the next to highest step on the progression. The highest step on all *GOLD*® progressions does not have behavioral anchors. Rather, it is to be used when a child's behavior exceeds expectations consistent with the behavioral anchors for the penultimate rating and is considered by the teacher to have advanced in their development beyond the scope of the progression.

It is important to note that these statistics are sample-specific. The sample used for this study contained a relatively small number of second-grade children and an even smaller number of third-grade children, and those children who were included were from a narrow range of classrooms. Therefore, these results may be an artifact of the limited sample of older children. It will be important to monitor these findings in future studies, especially after attaining larger and more diverse samples of second- and third-grade children. Similar findings of disorder at the top of the scale appear across the other domains and must be considered in light of these sample limitations.

Physical Scale

Three of the five progressions include a scale that ranges from 0 to 13, and two range from 0 to 15. There were no disordered category thresholds for three of the five progressions. Two progressions,

for Objectives 5 and 6, had disordered thresholds at the top of the progression. The observed sample averages for each response category advanced as expected across the rating scales with no disordered averages for three of the progressions. There were, however, disordered averages at the top of the scale for three of the progressions (Objectives 6, 7a, and 7b).

Language Scale

Six of the eight progressions include a scale that ranges from 0 to 15, one ranges from 0 to 11, and another ranges from 0 to 13. There were no disordered category thresholds for three of the eight progressions. Five of the progressions (Objectives 8a, 9a, 9b, 9c, and 10b) had disordered thresholds at the top of the scale. The observed sample averages for each response category generally advanced as expected across the rating scales with no disordered steps for two of the eight progressions. However, for six progressions (Objectives 8a, 9a, 9b, 9d, 10a, and 10b), there were disordered averages at the top of the scale.

Cognitive Scale

Six of the 10 progressions include a scale that ranges from 0 to 15, and four range from 0 to 13. There were no disordered category thresholds for 8 of the 10 progressions. Two progressions (Objectives 14a and 14b) had disordered thresholds at the top of the scale. The observed sample averages for each response category advanced as expected across the rating scales with no disordered steps for 6 of the 10 progressions. However, for four progressions (Objectives 11a, 11e, 12b, and 14b), there were disordered averages at the top of the scale.

Literacy Scale

Seven of the 16 progressions include a scale that ranges from 0 to 9, three range from 0 to 11, five range from 0 to 15, and one ranges from 0 to 19. For the literacy scale, there were issues with the category thresholds for the majority of progressions. For 10 of the 16 developmental progressions, there were disordered thresholds at the top of the scale. The observed sample averages for each response category generally advanced as expected across the rating scales. However, 4 of the 16 developmental progressions (Objectives 15c, 17a, 18a, and 18d) had disordered averages at the top of the scale.

Mathematics Scale

Six of the 12 progressions include a scale that ranges from 0 to 15, four range from 0 to 9, one ranges from 0 to 11, and one ranges from 0 to 13. For the mathematics scale, there were no issues with the category thresholds for 11 of the 12 progressions. One progression (Objective 20d) had disordered thresholds at the top of the scale. There were issues with disordered sample averages for the response categories at the top of the rating scales for 6 of the 12 developmental progressions (Objectives 20c, 20d, 20e, 20f, 21b, and 23).

These results suggest potential issues with assessing older children on some of the progressions. As already noted, however, it may be that these results are purely an artifact of the limited sample of older children. If they were to hold in a broader sample, however, they would indicate several avenues for additional investigation. The training and implementation processes for teachers of older children may need to be evaluated and monitored. The *GOLD*® approach is likely novel to many teachers in the early elementary grades. It is certainly plausible that second- and third-grade teachers need more support as they collect, interpret, and analyze their documentation. Alternatively, the utility and appropriateness of the behavioral anchors at the top of many of the progressions may require more in-depth examination. Ultimately, the need for additional analysis of larger and more diverse samples of children in the early elementary grades is indicated.

Progression Difficulty Measures

The Rasch measurement model estimates the difficulty level for each of the progressions within a given domain of development. These estimated difficulty levels are translated into locations on the same metric as the child ability estimates. Individual progressions can then be compared according to their relative difficulty levels and evaluated and interpreted as a developmental pathway from the least to most developmentally advanced. It is expected that most children will develop the skills and abilities associated with progressions located lower on the difficulty metric earlier than they will develop those associated with progressions located higher on the metric. For all six domains, the progression location hierarchy appeared to be generally consistent with the expected developmental trajectory for typically developing children. Hierarchies were also consistent with findings from previous studies. Tables 4 through 9 list the difficulty estimates from highest to lowest, along with the standard errors for these estimates and the associated fit statistics. These estimates were developed for the winter assessment period.

For the social–emotional scale, Objective 2a, "Forms relationships with adults," was the easiest skill progression (-2.80) by a large margin. Objective 3a, "Balances needs and rights of self and others," was the most difficult (1.29). The range of difficulties (-2.80 to 1.29) and rating scale threshold locations (-12.74 to 12.73) was wide enough to cover the full range of ability levels in the sample.

For the physical scale, Objective 7a, "Uses fingers and hands," was the easiest skill progression (-1.03), followed closely by Objective 4, "Demonstrates traveling skills." Objective 7b, "Uses writing and drawing tools," was the most difficult (1.27) by a wide margin. The range of difficulties (-1.03 to 1.27) and rating scale threshold locations (-16.27 to 11.32) was wide to cover the full range of ability levels in the sample .

For the language scale, Objective 8b, "Follows directions," was the easiest skill progression (-2.80) by a very wide margin. Objective 9d, "Tells about another time or place," was the most difficult (1.06). The range of difficulties (-2.80 to 1.06) and rating scale threshold locations (-15.97 to 11.14) was wide enough to cover the full range of ability levels in the sample.

For the cognitive scale, Objective 11b, "Persists," was the easiest skill progression (-1.60). Objective 12b, "Makes connections," was the most difficult (0.97), followed closely by Objective 11e, "Shows flexibility and inventiveness in thinking." The range of difficulties (-1.60 to 0.97) and rating scale threshold locations (-17.73 to 13.83) was wide enough to cover the full range of ability levels in the sample.

For the literacy scale, Objective 19a, "Writes name," was the easiest skill progression (-2.88). Objective 18d, "Uses context clues to read and comprehend texts," was the most difficult (2.99), followed closely by Objective 19c, "Writes using conventions." The range of difficulties (-2.88 to 2.99) and rating scale threshold locations (-14.34 to 9.27) was wide enough to cover the full range of ability levels in the sample.

For the mathematics scale, Objective 21a, "Understands spatial relationships," was the easiest skill progression (-4.11) by a very wide margin. Objective 20e, "Applies properties of mathematical operations and relationships," was the most difficult (2.37), followed closely by Objective 20d, "Understands and uses place value and base ten." The range of difficulties (-4.11 to 2.37) and rating scale threshold locations (-9.30 to 17.63) was wide enough to cover the full range of ability levels in the sample.

In summary, the developmental pathway for each scale indicates a progression from the easiest to the most difficult skills that aligns with expectations from developmental theory. In addition, the range of difficulties for each scale is the widest that has been observed across various validation studies. This finding suggests that teachers are able to use *GOLD*® to separate children according to their analysis of the evidence they collect.

Reliability

Reliability refers to how consistently a particular construct is measured by an assessment. This consistency can be measured by examining the internal consistency across progressions within a scale; the level of agreement between raters; or the extent to which scores for a child replicate across different assessments, testing situations, and/or time points. Reliability is considered an important psychometric characteristic of the information an assessment provides and is a necessary precondition for validity.

Reliability was evaluated using Cronbach's alpha measure of internal consistency. The simplest way to understand Cronbach's alpha is to start with the concept of split-half reliability. Split-half estimates of reliability involve splitting a test into two half-length versions, getting the correlation between these two halves, and then using a formula to correct for attenuation due to the shortened length. The purpose of the split-half reliability procedure is to provide an estimate similar to an equivalent-forms reliability when only a single form is available. In the case of *GOLD*®, this would involve dividing each domain into two subdomains, each with half of the objectives/dimensions, and then estimating the split-half reliability based on how well these two subdomains correlate. Cronbach's alpha provides a way to estimate this value without having to choose any one particular split; it gives the average of all possible split-half reliability estimates (Cronbach, 1951; Warrens, 2015). Alpha values above .70 are generally considered acceptable while values over .80 are good (George and Mallery, 2003).

Table 3 contains the reliability coefficients for each scale at each color band. Across all domains of development and all age-groups, alpha reliability estimates were at least 0.87 and reached as high as 0.98 for most domains in third grade. Note that reliabilities are not estimated for the literacy and math domains in the Birth to 1 year age-group. This is because, while these domains can be measured with *GOLD*® at this age-group, the vast majority of children fall into the "Not Yet" category on these dimensions and thus internal consistency reliability estimates are not informative.

Reliability estimates generally increased with age, which is to be expected, as older children can be scored on a larger number of dimensions within each domain and also display more variability. Other factors being equal, more variability yields higher reliability. The physical domain had lower internal consistency reliability than the other domains. This is due to the relatively low correlation between gross-motor skills (Objectives 4, 5, and 6) and fine-motor skills (Objectives 7a and 7b) as well as the relatively small number of dimensions in this domain. While the internal consistency of this domain is lower, test-retest and inter-rater reliabilities should be the same as other domains.

Inter-Rater Reliability

Inter-rater reliability on a subset of *GOLD*® progressions was investigated through a study conducted on Washington State kindergarten teachers (Joseph et al., 2020). Researchers developed a portfolio of videos showing four students displaying abilities related to 36 *GOLD*® progressions. This design was intentionally chosen to match the approach of the *GOLD*® Interrater Reliability Certification process at that time. Fifty-four (54) teachers rated each child on all progressions, and these ratings were compared to scores assigned by a highly trained master rater.

The percentage of exact plus adjacent agreement ranged from 68.4% to 84.4%. Agreement was higher for social–emotional, physical, and language progressions. Lower levels of agreement were found for five specific progressions (Objetives 11c, 13, 18a, 18c, and 20b). It is probable that the recorded videos provided in this study did not provide adequate evidence for teachers to rate these dimensions with a high degree of accuracy. Notably, the study also found that achieving *GOLD*® Interrater Reliability Certification had a significant positive effect on teacher's agreement with the master coding (Josept et al., 2019) .

Reliability Analyses by Subgroup

Standard 2.11 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014, p. 45) states, "Test publishers should provide estimates of reliability/precision as soon as feasible for each relevant subgroup for which the test is recommended." In this section, we provide coefficient alpha reliability estimates for gender, race/ethnicity, language, disability status, and free-reduced lunch status subgroups.

Tables 22–26 contain comparisons of reliability coefficients by gender, race/ethnicity, language, disability status, and free-reduced lunch (FRL) status. All categories except FRL show results for the first six color bands. Free-reduced lunch information was only available at Preschool 3 or higher, so for this comparison, only coefficients for Preschool 3, Pre-K 4, and Kindergarten could be estimated. Within gender, all reliabilities were .84 or greater, all comparisons were within .03 of each other, and most were within .01. For race/ethnicity, language, and disability status, the values for all subgroups were strong (.82–.97). With a few exceptions, most comparisons were within .02. Students on free-reduced lunch showed reliabilities of .86 or higher on all domains, and most comparisons were within .01.

Validity

Concurrent Validity

The concurrent validity of *GOLD*® was investigated in a 2013 study by the American Institutes for Research (AIR). Moderate correlations were found between social–emotional scaled scores and all the Preschool and Kindergarten Behavior Scales (PKBS) (Merril, 2003) subscales (r = .428 to r = .523), with the exception of the PKBS Internalizing and Externalizing Problem Behaviors, which showed weak correlations. The PKBS Social Interaction score was most strongly associated with the *GOLD*® cognitive scaled score (r = .541). Scores on the Persistence and Motivation subscales of the Preschool Learning Behaviors Scale (PLBS) (McDermott, Leigh, and Perry 2002) were moderately correlated with *GOLD*® cognitive scaled scores (range r = .428 to r = .486). The Total PKBS score was moderately associated with all *GOLD*® scaled scores (r = .426 to r = .507). Scores on the Peabody Picture Vocabulary Test (Dunn & Dunn, 2007) were moderately associated with *GOLD*® language, literacy, mathematics, and cognitive scaled scores (r = .436 to r = .483). Scores on the Head-Toes-Knees-Shoulders task (HTKS) (Ponitz, McClelland, Matthews, & Morrison, 2009) was moderately associated with the *GOLD*® language, literacy, and mathematics scaled scores (r = .356 to r = .389).

Scores on the Woodcock Johnson III (WJ-III) Tests of Achievement (Woodcock, McGrew, & Mather, 2007) Letter-Word Identification, Word Attack, and Understanding Directions were moderately associated with *GOLD*® literacy scaled scores (range r = .372 to r = .450). The Quantitative Concepts score was moderately associated with all the *GOLD*® scaled scores (r = .399 to r = .522) and, as expected, the highest correlation was with *GOLD*® mathematics.

These associations with the WJ-III were replicated more recently by Miller-Bains et al. (2017). This study found a strong correlation between WJ Basic Reading and *GOLD*® literacy (r = .675), moderate correlations between WJ Picture Vocabulary and *GOLD*® literacy and language (r = .472 and = .400), and a strong correlation between WJ Math Reasoning and *GOLD*® mathematics (r = .558).

Validity Evidence Based on Examination of Subgroups

Given the multi-tiered nature of the complex response process involved in the use of authentic formative assessment measures such as *GOLD*® and the multiple possibilities for teachers to introduce construct-irrelevant variance that can impact the fairness of assessment scores, authentic formative assessments such as *GOLD*® qualify as what Engelhard (2002; Engelhard & Wind, 2018) has called "rater-mediated assessments." Therefore, *GOLD*® assessment scores are mediated by, or emerge through, the judgement of the teacher (Engelhard & Wind, 2018). Within this process, teachers

analyze and interpret evidence of child progress elicited by the regular instructional process. Then the teacher, not the child, responds to the indicators of ability contained within the developmental progressions by making ratings. However, interpretation of the resulting assessment scores has consequences for the child, not the teacher. Ideally, the scores, when properly interpreted, lead to decisions about instructional goals, strategies, and support provided to the child. Therefore, the scores yielded by measures such as *GOLD*® lead to meaningful, useful, and actionable conclusions for all children in the classroom only when teachers can apply the rating scales in a reliable and fair manner across all subgroups of children.

Unlike direct standardized assessments, authentic observational assessments rely heavily on the teacher's knowledge of child development. The teacher takes an active role in all aspects of the assessment process, including observations across contexts, selection of evidence and artifacts, analysis and interpretation of evidence, and using professional judgment to make placements on developmental progressions (Heritage, 2013; Lambert, 2020).

Misapplication of the formative assessment process in any of these areas creates the potential for construct-irrelevant variance to influence assessment scores. *GOLD*® offers inter-rater reliability certification to mitigate some sources of rater error. However, any rater-mediated assessment requires ongoing iterative research to study rater effects and their influence on the response process to ensure the assessment is fair to all subgroups, including children with disabilities (AERA, APA, & NCME, 2014).

Standard 3.0 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014, p. 63) makes it clear that "All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population." The Standards identify the following potential sources of construct-irrelevant variance in assessment scores: a.) assessment content, b.) assessment context, c.) response process, and d.) opportunity to learn. Validity evidence based on the integrity and fairness of the response process is one of the most important components of an argument to support the meaningfulness, usefulness, and appropriateness of the interpretations of assessment scores (Messick, 1995; AERA, APA, & NCME, 2014).

Furthermore, the standards specifically require evidence that assessment scores can be interpreted in the same way, or have the same essential meaning, across all subgroups of children (AERA, APA, & NCME, 2014). The Standards, when applied to rater-mediated assessments for young children, indicate that rater effects on the response process should be minimized so that an assessment system is fair to all subgroups of children. Therefore, variability in assessment scores should correspond to variability in true differences in child ability and be free of the influence of any other teacher or child characteristics.

However, teachers can introduce construct-irrelevant variance into a set of scores derived from an authentic formative assessment by applying the complex response process in different ways across subgroups of children. This can result in a lack of fairness within the assessment process, and fairness in testing is a fundamental principle of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The *Standards* define fairness in terms of the absence of measurement bias. Any introduction of construct-irrelevant variance that results in assessment scores taking on different meaning for various subgroups of children is evidence of a lack of fairness. The *Standards* identify differential item functioning (DIF) analyses as a viable strategy for contributing to an argument for the fairness of the information provided by assessment scores under specific conditions of use (AERA, APA, & NCME, 2014, p. 51).

Differential Item Functioning Overview

Differential item functioning (DIF) analysis provides validity evidence as to whether subgroup differences exist in the application and interpretation of assessment outcomes by examining subgroup differences in item difficulty. Measurement invariance theory suggests students with similar abilities from different subgroups should demonstrate similar observed scores (Wu et al., 2007). DIF analysis can provide one type of evidence that the response process has been applied consistently across subgroups. "DIF centers on identifying whether examinee responses to particular test items or sets of test items are linked systematically to the personal characteristics (e.g., gender or ethnicity) of the examinees and are otherwise unrelated to the test's central construct" (Osterlind and Everson, 2009, p.3). In other words, if evidence of DIF exists for specific items or sets of items for a particular subgroup, such evidence would indicate that those items need further investigation to ensure meaningful, appropriate, and useful interpretation of test scores.

To examine DIF, researchers match members of the subgroup of interest to members of the reference group by their overall ability levels and then compare item responses across these groups. Evidence of DIF would suggest further research is needed to investigate potential test bias or concerns with rater characteristics and behaviors. Evidence of DIF also points to the need to further investigate whether the measure in question may actually measure different constructs for various subgroups.

Specifically, DIF analyses can address these issues by providing evidence that the response process does not (dis)advantage any particular subgroup of children. DIF analyses separate performance differences on specific items from individual differences on the measured latent construct. Focal and reference categories of examinees are identified to investigate DIF. Researchers examine the focal group to gather evidence to demonstrate that the assessment process is fair to the members of that particular subgroup. The historically advantaged group is typically the reference group, with the presumption that the assessment process is generally fair for them. Researchers compare the focal and reference groups by examining their item responses after first matching them on their overall ability estimates.

Respondents are, therefore, matched based on the estimated quantity of the latent construct under investigation they are presumed to possess (Walker, 2011). In a Rasch modeling context, DIF can be thought of as differences between members of the focal and reference groups in the conditional probability of a particular item response.

Non-Rasch DIF procedures investigate the focal and reference groups for interactions between sample characteristics and item responses while holding overall ability levels of the test takers constant. The analytical process estimates overall ability for each participant from total performance across the full item set. Therefore, such DIF statistical procedures answer the following research question: "Do test takers in the focal and reference groups who have the same overall ability levels respond in similar ways to each item?" Evidence of DIF can lead to further investigation of item content, item interpretation by respondents, rater characteristics and behavior, construct validity of the items and measures, and potential test bias.

From a test validity perspective, item functioning needs to be investigated for possible interactions with characteristics of subgroups of test takers to ensure that the test is measuring the same construct of interest, and only the construct of interest, in the same way for all subgroups (Badia, X., Prieto, L., and Linacre, J. M., 2002). This broad principle of validity extends to teacher ratings and authentic formative assessments and is an important part of the validity argument to support the use of the *GOLD*® assessment system. Therefore, this section focuses on existing validity evidence for the use of *GOLD*® with various subgroups of children.

Summary of Existing *GOLD*® Subgroup Validity Evidence

Kim, Lambert, and Burts (2013) conducted a study assessing the validity of the information *GOLD*® provides when used with English-language learners (the conventional term at that time) and children with disabilities. The study examined if, when teachers are aware of a child's English-language learner (ELL) or disability status, they assign the same scores to these subgroups of children as they do to children of similar overall ability who are not ELLs or who do not have a disability. Specifically, assessment information was collected on 3-, 4-, and 5-year-old children at the fall (n = 79,324), winter (n = 132,693), and spring (n = 50,558) checkpoints. DIF analyses indicated that, in general, teachers' ratings were similar for children of similar abilities, regardless of their subgroup membership. For children with disabilities, results indicated that most progressions had negligible DIF; however, four progressions exhibited either intermediate or large DIF consistently favoring children without disabilities. Those progressions pertain to children's expressive language abilities and ability to engage in conversation (Kim et al., 2013). For ELLs, again, results indicated neglible DIF for most progressions, except progression 9c, "Uses conventional grammar" (Kim et al., 2013): this progression consistently favored non-ELLs.

Similarly, in a DIF study of 4-year-old dual-language learners (DLLs), the conventional term at that time, and native English speakers assessed using *GOLD*® (n = 133,732), researchers found that almost all progressions functioned similarly when comparing three distinct latent subgroups of DLL children (Kim, Lambert, Durham, & Burt, 2018). Just one of the *GOLD*® developmental progressions displayed non-negligible DIF in ways that favored native English-speakers: Objective 9c, "Uses conventional grammar."

Associations of teacher ratings with child demographics (namely age, gender, disability status, and ELL status) and classroom composition characteristics (namely class mean age and percentage ELLs, children with disabilities, and males) were examined with a sample of 21,592 children ages 12–59 months old (Lambert, Kim, & Burts, 2014). Using three-level growth curve modeling, findings indicated that teachers' *GOLD*® ratings were associated with the child and classroom characteristics in the anticipated directions for all subgroups of interest. Children with disabilities began the year with lower scores on *GOLD*® relative to those of their typically developing peers and grew more slowly throughout the year. Girls demonstrated higher scores than boys in some developmental domains. ELLs had lower scores at the beginning of the year but exhibited somewhat faster growth rates than native English speakers.

Current Differential Item Functioning Analyses

A new round of DIF analyses was undertaken to investigate whether teachers displayed different patterns of item response when rating specific subgroups of children. In this study, "item response" was defined as placements made by teachers on *GOLD*® developmental progressions. Given that *GOLD*® is an authentic formative assessment, "item response" does not refer to direct answers from children. Rather, it refers to the placements on *GOLD*® developmental progressions assigned to each child by teachers based on evidence produced by the child in the course of normal instructional activities. Specifically, different patterns of response were defined in terms of DIF. DIF focuses on whether test items function in different ways when teachers rate different subgroups of children. Ideally, item functioning is invariant to the specific aspects of the children assessed, such as class demographics, classroom environment, or teacher characteristics that are considered irrelevant to measuring the underlying ability or latent construct that is the focus of the assessment.

In order to interpret the findings from this study properly, it is important to recognize exactly what information DIF analyses can and cannot provide. In the context of this study, DIF is present when members of different subgroups of children, for a particular *GOLD*® developmental progression, have different probabilities of placement at specific levels on the rating scale structure, after the overall developmental level of each child on the domain in question has been accounted for.

Furthermore, DIF in the context of this study can suggest the possibility that teachers may be biased in their ratings of children of the focal group relative to how they rate the reference group. Evidence of DIF can suggest a need to investigate potential biases but does not guarantee the presence of item or rater bias.

All previous *GOLD*® DIF studies utilized data yielded by an older version of *GOLD*®, the birth to kindergarten version of the measure. The *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 2014) indicate that if revisions influence the interpretation of scores, assessment developers have the responsibility to offer new validity evidence in support of the use of the new version of a measure. Specifically, Standards 4.24 and 4.25 require assessment developers to provide users with guidelines concerning the appropriate uses of, and the comparability of scores from, earlier and revised versions of a measure. Therefore, researchers and the developers of *GOLD*® should provide users with a strong validity argument for the use of the revised *GOLD*® measure relative to its intended purposes. This phase of research sought to extend previous studies and establish further validity evidence for the birth to third grade version of *GOLD*®.

Specifically, this research sought to examine potential evidence for DIF in the placements teachers make on *GOLD*® developmental progressions based on gender, primary language spoken in the home, race/ethnicity, and IEP status. This research examined teachers' response processes in the placement of children on *GOLD*® developmental progressions to determine whether teachers display different patterns of item response when rating students identified in specific subgroups. This type of evidence can enhance teacher confidence in the utility of the information an assessment provides. When teachers and families trust the validity of an assessment, they are more likely to utilize outcomes for instructional planning. Meaningful application of information garnered from authentic formative assessments leads to informed classroom practices and individualized instruction that facilitates students' successful acquisition of developmental milestones.

Rasch modeling was used to produce composite interval level scaled scores for each domain of development, item difficulties, and child ability estimates that are useful for statistical comparisons in general and testing for DIF in particular. Furthermore, this study sought to further the validity argument for the use of *GOLD*® with young children by extending and updating previous DIF research through a focus on data collected using the latest version of the measure.

Methods

Assessment records from the 2018–2019 academic year for students with data from fall, winter, and spring were used to study DIF by gender, primary language, and race/ethnicity. The sample used to examine DIF by IEP status was selected from the entire national database of *GOLD*® users across the U.S. Children were included in the sample if they met the following criteria: a) the program they attend used *GOLD*® to document the developmental status of preschool or prekindergarten children during the 2020-2021 academic year to inform instructional planning and communicate with stakeholders; b) the program they attend used *GOLD*® for Office of Special Education Programs (OSEP) reporting purposes; and c) the program included information about Individualized Family Service Plans (IFSP), Individual Education Programs (IEP), and applicable information about entry and exit from early intervention services (Part C of the Individuals with Disabilities Education Act) and services for school-aged children (Part B of the Individuals with Disabilities Education Act) within the *GOLD*® data reporting system. The children in the sample (n = 105,920) attended publicly funded preschool and prekindergarten classrooms within programs that met these criteria. The qualifying programs were in the following states: Colorado, Nebraska, Nevada, New Hampshire, Arizona, and Louisiana.

It is important to note the importance of focusing on programs that met these criteria for this study. These programs were more likely to serve children with disabilities in inclusive classrooms and to serve a higher percentage of children with disabilities than typical publicly funded early childhood programs in the U.S. In this sample, 21.1% of children had an IEP, as compared to 13.2% for the national Head Start program (Early Childhood Learning and Knowledge Center, 2020). In addition, 1.9% of the children also had an IFSP prior to having an IEP. Therefore, teachers in these programs had more experience assessing children with disabilities than would many teachers who serve fewer children with IEPs. Most importantly, the information in the *GOLD*® system about which children have an IEP was more likely to be accurate in programs that use *GOLD*® for OSEP reporting. Programs that do not use *GOLD*® for OSEP reporting are more likely to leave this information unreported. Given the focus of the analyses was comparisons between children with and without an IEP, this criterion was an essential part of the sampling process.

Demographic characteristics of the DIF sample are presented in Table 45. Most of the children attended 4-year-old prekindergarten classrooms (65.5%). The remaining children attended 3-year-old preschool classrooms (35.5%). There were slightly more boys in the sample than girls (boys=46.8%, girls=53.2%). The majority of children came from homes in which English was the primary language spoken (87.9%). Spanish was the primary language spoken at home for 10.2% of the children, and some other language was spoken in the homes of the remaining children (1.9%). Dual-language learners comprised 13.8% of the sample. The race/ethnicity of the children in the sample was comprised of White (42.3%), Black (23.1%), Asian (2.2%), Native American/ Pacific Islander (2.1%), Hispanic (27.1%), and two or more races (3.2%). Children who qualified for the National School Lunch Program (free or reduced-price lunch), as reported by their teacher, comprised 24.0% of the children.

Data Analysis

First, interval level measures were created for each of the *GOLD*® domains of development using the Rasch partial credit model (PCM; Masters, 1982), as operationalized through the Winsteps software (version 4.6.2.1) (Linacre, 2020). Specifically, the resulting measures were created using the Linacre partial credit model, an extension of the Rasch rating scale model (Andrich, 1978) and Masters partial credit model (Masters, 1982). This step allows researchers to access scaled scores for each participating child that were measured on the same interval scale. This method also provided invaluable diagnostic information about the fit of the data yielded by each progression to the Rasch measurement model. A separate Rasch analysis was conducted for each of the six domains of development. The Rating Scale (RSM) (Andrich, 1978, Bond & Fox, 2001) and the PCM are the two most widely used Rasch models for polytomous response data. The RCM is useful when all items share the same rating scale. In cases where each item has its own rating scale structure, the PCM is the appropriate model to apply. The PCM, rather than the RCM, was chosen because the *GOLD*® progressions do not share the same rating scales (i.e., the number of rating scale categories and category labels varies across progressions).

The following procedures were used to test for the presence of DIF and to examine the magnitude of DIF. First, the difficulty level for each progression was estimated for both the focal and reference groups. The DIF contrast statistic, the difference between the paired Rasch difficulty estimates for each progression, was calculated as the simple difference between the difficulty estimates for the focal and reference groups. Next, Rasch-Welch t tests were examined for statistical significance. These statistics report whether item difficulty estimates for the two groups, across each of the progressions, are the same except for measurement error. These tests evaluate a null hypothesis that the DIF contrast value is zero against an alternative hypothesis that the DIF contrast is not equal to zero. Next, the Mantel-Haenszel X^2 statistics produced by the Winsteps software package (Linacre, 2020) were also used to examine evidence of potential DIF. These tests examine a null hypothesis of no DIF by producing a probability of obtaining differences between the focal and reference groups as large as or larger than those obtained, given that there is no DIF. The groups

are stratified into matching ability levels, and their relative performance on each progression is quantified. Overall ability estimates for each respondent are estimated based on the total scores on the measure. An alpha level of .05 was used for all comparisons.

In addition to statistical significance testing, examination of the magnitude of the DIF contrast, or difference in item difficulty estimates between the focal and reference groups, is critical given the sensitivity of both the Rasch-Welch *t* test and the Mantel-Haenszel χ^2 to small differences. This is especially important in large-sample size studies like this one, where statistical significance is easy to obtain even when the observed differences do not have practical implications. Magnitude of the DIF contrast was determined according to the criteria set forth by Zwick, Thayer, and Lewis (1999). If both the t and χ^2 statistics were statistically significant and the magnitude of the DIF contrast was less than .43, the DIF magnitude was considered negligible. If both the *t* and χ^2 statistics were statistically significant and the magnitude of the DIF contrast was greater than or equal to .43 and less than .64, the DIF magnitude was considered intermediate. If both the *t* and χ^2 statistics were statistically significant and the magnitude of the DIF contrast was greater than .64, the DIF magnitude was considered large. To further aid interpretation of subgroup comparisons, progressions with estimated difficulty measures in specific ranges, by subgroup, were classified as follows: below −.5 were considered easy, −.5 to .5 were considered average, and those with values above .5 were considered difficult. For each domain of development, the average difficulty was set to a logit value of zero.

Specifically, this phase of research posed the following questions.

1. Is there evidence of DIF regarding the placements made by teachers on *GOLD*® developmental progressions for male versus female children?
2. Is there evidence of DIF regarding the placements made by teachers on *GOLD*® developmental progressions for children living in homes where English versus Spanish is the primary language spoken?
3. Is there evidence of DIF regarding the placements made by teachers on *GOLD*® developmental progressions for white versus non-white children?
4. Is there evidence of DIF regarding the placements made by teachers on *GOLD*® developmental progressions for children without an IEP versus children with an IEP?

For analyses of DIF by gender, male children were the focal group and female students were the reference group. This decision was made because the overwhelming majority of teachers are female. For DIF analyses by primary language, children living in homes where Spanish is the primary language were the focal group and children living in homes where English is the primary language were the reference group. This decision was made because most early childhood teachers in the United States are native English speakers and English is the most frequently spoken language. For DIF by race/ethnicity, non-white children were the focal group and white children were the reference group. White children were the reference group because the majority of the early childhood workforce consists of white teachers. For DIF by IEP status, children without an IEP were the reference group and children with an IEP were the focal group.

DIF Analyses by Gender

Social–Emotional

For five of the nine progressions within the social–emotional domain (Objectives 1a, 1c, 2c, 3a, and 3b), neither the Rasch-Welch t nor the Mantel-Haenszel χ^2 statistics were statistically significant. For one of the progressions within the social–emotional domain (Objective 2b), only the Mantel-Haenszel χ^2 statistic was statistically significant ($p < .05$). For the remaining three progressions within the social–emotional domain (Objectives 1b, 2a, and 2d), both the Rasch-Welch t and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .001$). The estimated difficulties for

males and females were identical for six of the nine progressions (Objectives 1a, 1c, 2b, 2c, 3a, and 3b). The differences in estimated difficulties for all progressions were very small (.05 to .11). As seen in Table 27, the difficulty classifications (Easy, Average, or Difficult) were the same for both males and females across all nine progressions. Similarly, the DIF magnitudes were considered negligible for all nine progressions.

Physical

For all five progressions within the physical domain (Objectives 4, 5, 6, 7a, and 7b), both the Rasch-Welch t and the Mantel-Haenszel χ^2 statistics were statistically significant (p <.001). The differences in estimated difficulties for all five progressions were small (-.23 to .25). As seen in Table 28, the difficulty classifications (Easy, Average, or Difficult) were the same for both males and females across four of the five progressions. One progression (Objective 6), met the criteria for difficult for females and the criteria for average for males. This progression focuses on gross-motor manipulation skills, and these results suggest it may be slightly more difficult for teachers to move girls to the next level on the progression as compared to boys. However, the DIF magnitudes were considered negligible for all five progressions.

Language

Across four of the eight progressions within the language domain (Objectives 9a, 9c, 10a, and 10b), neither the Rasch-Welch t nor the Mantel-Haenszel χ^2 statistics were statistically significant. For one of the progressions within the language domain (Objective 8b), only the Mantel-Haenszel χ^2 statistics were statistically significant (p <.05). For the remaining three progressions within the language domain (Objectives 8a, 9b, and 9d), both the Rasch-Welch t and Mantel-Haenszel χ^2 statistics were statistically significant (p <.05). Across four of the eight progressions (Objectives 9a, 9c, 10a, and 10b), the estimated difficulties for males and females were identical. The differences in estimated difficulties for all progressions were very small (-.05 to .05). As seen in Table 29, the difficulty classifications (Easy, Average, or Difficult) were the same for both males and females across all eight progressions. Similarly, the DIF magnitudes were considered negligible for all eight progressions.

Cognitive

Across 4 of the 10 progressions within the cognitive domain (Objectives 11c, 12a, 12b, and 14a), neither the Rasch-Welch t nor the Mantel-Haenszel χ^2 statistics were statistically significant. For the remaining six progressions within the cognitive domain (Objectives 11a, 11b, 11d, 11e, 13, and 14b), both the Rasch-Welch t and Mantel-Haenszel χ^2 statistics were statistically significant (p <.05). The estimated difficulties for males and females were identical across 4 of the 10 progressions (Objectives 11c, 12a, 12b, and 14a). The differences in estimated difficulties for all progressions were very small (-.13 to .11). As seen in Table 30, the difficulty classifications (Easy, Average, or Difficult) were the same for both males and females across all 10 progressions. Similarly, the DIF magnitudes were considered negligible for all 10 progressions.

Literacy

For 6 of the 16 progressions within the literacy domain (Objectives 15d, 17a, 17b, 18a, 19a, and 19c), neither the Rasch-Welch t nor the Mantel-Haenszel χ^2 statistics were statistically significant. Across four of the progressions within the literacy domain (Objectives 15b, 15c, 18b, and 18c), only the Mantel-Haenszel χ^2 statistic was statistically significant (p <.05). For one of the progressions (Objective 18e), only the Rasch-Welch t statistic was statistically significant (p <.05). For the remaining five progressions within the literacy domain (Objectives 15a, 16a, 16b, 18d, and 19b), both the Rasch-Welch t and Mantel-Haenszel χ^2 statistics were statistically significant (p <.05). The estimated difficulties for males and females were identical for 9 of the 16 progressions (15b, 15c, 17a, 17b, 18a, 18b, 18c, 19a, and 19c). The differences in estimated difficulties for all progressions were very small (-.05 to .12). As seen in Table 31, the difficulty classifications (Easy, Average, or Difficult) were the same for both males and females across all 16 progressions. Similarly, the DIF magnitudes were considered negligible for all 16 progressions.

Mathematics

For 7 of the 12 progressions within the mathematics domain (Objectives 20a, 20b, 21b, 22a, 22b, 22c, and 23), neither the Rasch-Welch t nor the Mantel-Haenszel χ^2 statistics were statistically significant. For the remaining five progressions within the mathematics domain (Objectives 20c, 20d, 20e, 20f, and 21a), both the Rasch-Welch t and Mantel-Haenszel χ^2 statistics were statistically significant (p <.05). The estimated difficulties for males and females were identical for 7 of the 12 progressions (Objectives 20a, 20b, 21b, 22a, 22b, 22c, and 23). The differences in estimated difficulties for all progressions were very small (-.07 to .07). As seen in Table 32, the difficulty classifications (Easy, Average, or Difficult) were the same for both males and females across all 12 progressions. Similarly, the DIF magnitudes were considered negligible for all 12 progressions.

DIF Analyses by Primary Language

Social–Emotional

For three of the nine progressions within the social–emotional domain (Objectives 1a, 1b, and 2b), neither the Rasch-Welch t nor the Mantel-Haenszel χ^2 statistics were statistically significant. For the remaining six progressions within the social–emotional domain (Objectives 1c, 2a, 2c, 2d, 3a, and 3b), both the Rasch-Welch t and Mantel-Haenszel χ^2 statistics were statistically significant (p <.05). The estimated difficulties for children from homes where English is the primary language (EPL) and children from homes where Spanish is the primary language (SPL) were identical for Objective 1b. The differences in estimated difficulties for all progressions were very small (-.15 to .18). As seen in Table 33, the difficulty classifications (Easy, Average, or Difficult) were the same for both EPL and SPL children across eight of the nine progressions. The only exception to this finding was for Objective 2b. The estimated difficulty for EPL children was .49, which is at the top of the Average range, and the estimated difficulty for SPL children was .51, which is at the bottom of the Difficult range, resulting in a very small difference. The DIF magnitudes were considered negligible for all nine progressions.

Physical

For one progression within the physical domain (Objective 4), neither the Rasch-Welch t nor the Mantel-Haenszel χ^2 statistics were statistically significant. For one progression (Objective 7a), only the Mantel-Haenszel χ^2 statistics was statistically significant (p <.05). For the remaining three progressions within the physical domain (5, 6, and 7.b), both the Rasch-Welch t and the Mantel-Haenszel χ^2 statistics were statistically significant (p <.05). The differences in estimated difficulties for all five progressions were very small (-.10 to .18). As seen in Table 34, the difficulty classifications (Easy, Average, or Difficult) were the same for both EPL and SPL children across all five progressions. Similarly, the DIF magnitudes were considered negligible for all five progressions.

Language

For three of the eight progressions within the language domain (Objectives 8a, 9a, and 10a), neither the Rasch-Welch t nor the Mantel-Haenszel χ^2 statistics were statistically significant. For one of the progressions within the language domain (Objectives 9d), only the Mantel-Haenszel χ^2 statistic was statistically significant (p < .05). For the remaining four progressions within the language domain (Objectives 8b, 9b, 9c, and 10a), both the Rasch-Welch t and Mantel-Haenszel χ^2 statistics were statistically significant (p < .001). The differences in estimated difficulties for all progressions were small (-.20 to .23). As seen in Table 35, the difficulty classifications (Easy, Average, or Difficult) were the same for both EPL and SPL children across all eight progressions. Similarly, the DIF magnitudes were considered negligible for all eight progressions.

Cognitive

For 2 of the 10 progressions within the cognitive domain (Objectives 13 and 14b), neither the Rasch-Welch t nor the Mantel-Haenszel χ^2 statistics were statistically significant. For 2 of the 10 progressions (Objectives 11e and 14a), only the Mantel-Haenszel χ^2 statistic was statistically

significant ($p < .05$). For the remaining six progressions within the cognitive domain (Objectives 11a, 11b, 11c, 11d, 12a, and 12b), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .05$). The differences in estimated difficulties for all progressions were small (-.10 to .27). As seen in Table 36, the difficulty classifications (Easy, Average, or Difficult) were the same for both EPL and SPL children across all 10 progressions. Similarly, the DIF magnitudes were considered negligible for all 10 progressions.

Literacy

For 3 of the 16 progressions within the literacy domain (Objectives 15c, 17b, and 18b), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For two of the progressions within the literacy domain (Objectives 15b and 18d), only the Rasch-Welch *t* statistic was statistically significant ($p < .05$). For the remaining 11 progressions within the literacy domain (Objectives 15a, 15d, 16a, 16b, 17a, 18a, 18c, 18e, 19a, 19b, and 19c), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .05$). The estimated difficulties for both groups were identical for 3 of the 16 progressions (Objectives 15c, 17b, and 18b). The differences in estimated difficulties for all progressions were small (-.12 to .22). As seen in Table 37, the difficulty classifications (Easy, Average, or Difficult) were the same for both EPL and SPL children across all 16 progressions. Similarly, the DIF magnitudes were considered negligible for all 16 progressions.

Mathematics

For 6 of the 12 progressions within the mathematics domain (Objectives 20a, 20c, 20d, 20f, 22c, and 23), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For the remaining six progressions within the mathematics domain (Objectives 20b, 20e, 21a, 21b, 22a, and 22b), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .01$). The estimated difficulties for EPL and SPL children were identical for 3 of the 12 progressions (Objectives 20a, 20c, and 23). The differences in estimated difficulties for all progressions were very small (-.10 to .13). As seen in Table 38, the difficulty classifications (Easy, Average, or Difficult) were the same for both EPL and SPL children across all 12 progressions. Similarly, the DIF magnitudes were negligible for all 12 progressions.

DIF Analyses by Race/Ethnicity

Social–Emotional

For one of the nine progressions within the social–emotional domain (Objective 2b), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For two of the nine progressions within the social–emotional domain (Objectives 1a and 2c), only the Mantel-Haenszel χ^2 statistics were statistically significant ($p < .01$). For the remaining six progressions within the social–emotional domain (Objectives 1b, 1c, 2a, 2d, 3a, and 3b), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .001$). The estimated difficulties for white and non-white children were identical for Objectives 2b and 2c. The differences in estimated difficulties for all progressions were small (-.21 to .14). As seen in Table 39, the difficulty classifications (Easy, Average, or Difficult) were the same for both white and non-white children across all nine progressions. Similarly, the DIF magnitudes were negligible for all nine progressions.

Physical

For two progressions within the physical domain (Objectives 4 and 7a), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For two progressions (Objectives 5 and 6), only the Mantel-Haenszel χ^2 statistics were statistically significant ($p < .05$). For Objective 7b, both the Rasch-Welch *t* and the Mantel-Haenszel χ^2 statistics were statistically significant ($p < .001$). The estimated difficulties for white and non-white children were identical across four of the five progressions (Objectives 4, 5, 6, and 7a). The differences in estimated difficulties for all five progressions were very small (-.06 to .00). As seen in Table 40, the difficulty classifications (Easy, Average, or Difficult) were the same for both white and non-white children across all five progressions. Similarly, the DIF magnitudes were negligible for all five progressions.

Language

For one of the eight progressions within the language domain (Objective 10a), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For the remaining seven progressions within the language domain (Objectives 8a, 8b, 9a, 9b, 9c, 9d, and 10b), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .05$). The estimated difficulties for white and non-white children were identical for one progression (Objective 10a). The differences in estimated difficulties for all progressions were very small (-.11 to .14). As seen in Table 41, the difficulty classifications (Easy, Average, or Difficult) were the same for both white and non-white children across all eight progressions. Similarly, the DIF magnitudes were negligible for all eight progressions.

Cognitive

For 6 of the 10 progressions within the cognitive domain (Objectives 11b, 11c, 11e, 12a, 12b, and 13), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For the remaining four progressions within the cognitive domain (Objectives 11a, 11d, 14a, and 14b), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .001$). The estimated difficulties for white and non-white children were identical for five progressions (Objectives 11b, 11e, 12a, 12b, and 13). The differences in estimated difficulties for all progressions were very small (-.13 to .10). As seen in Table 42, the difficulty classifications (Easy, Average, or Difficult) were the same for both white and non-white children across 9 of the 10 progressions. The only exception to this finding was for Objective 11d. The estimated difficulty for white children was .50, which is at the top of the Average range, and the estimated difficulty for non-white children was .63, which is at the bottom of the Difficult range, resulting in a very small difference. The DIF magnitudes were negligible for all 10 progressions.

Literacy

For 9 of the 16 progressions within the literacy domain (Objectives 15b, 15c, 16b, 17b, 18a, 18b, 18c, 19a, and 19b), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For the remaining seven progressions within the literacy domain (Objectives 15a, 15d, 16a, 17a, 18d, 18e, and 19c), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .001$). The estimated difficulties for white and non-white children were identical for 7 of the 16 progressions (Objectives 15c, 16b, 17b, 18b, 18c, 19a, and 19b). The differences in estimated difficulties for 14 of the 16 progressions were small (-.13 to .28). The only exceptions to this finding were for Objective 18d, where the estimated difference was .39, and Objective 19c, where the estimated difference was .41.

These differences are at the upper end of the negligible range. However, both progressions were very difficult for white and non-white children. As seen in Table 43, the difficulty classifications (Easy, Average, or Difficult) were the same for both white and non-white children across all 16 progressions. Similarly, the DIF magnitudes were considered negligible for all 16 progressions.

Mathematics

For 5 of the 12 progressions within the mathematics domain (Objectives 20a, 20c, 20d, 22c, and 23), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For one progression (Objective 20f), only the Rasch-Welch *t* statistic was statistically significant ($p < .05$). For the remaining six progressions within the mathematics domain (Objectives 20b, 20e, 21a, 21b, 22a, and 22b), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .01$). The estimated difficulties for white and non-white children were identical for 3 of the 12 progressions (Objectives 20a, 20b, and 20c). The differences in estimated difficulties for all progressions were small (-.09 to .25). As seen in Table 44, the difficulty classifications (Easy, Average, or Difficult) were the same for both white and non-white children across all 12 progressions. Similarly, the DIF magnitudes were considered negligible for all 12 progressions.

DIF Analyses by IEP Status

The demographic characteristics of the sample of children involved in the analyses to address DIF by IEP status are reported in Table 45. The sample largely reflected the population of young children who attend preschools in the United States. However, children with disabilities were somewhat overrepresented in the sample. Only programs that met specific criteria, discussed earlier in this manual, were selected to contribute to this sample. The criteria were designed to limit the sample to early childhood programs with trustworthy data regarding children with IEPs. The qualifying programs are more likely, therefore, to focus on and serve more children with identified special needs than most other programs.

Social–Emotional

The Rasch-Welch *t* was statistically significant for all nine progressions within the social–emotional domain. The Mantel-Haenszel χ^2 statistic was statistically significant for eight of the nine progressions. However, the magnitude of the DIF contrasts were negligible for all nine progressions. The differences in estimated difficulties were very small in absolute value (.03 to .22) for eight of the nine progressions. Only one progression, Objective 2a, yielded a DIF contrast statistic (.42) that was close to the upper end of the negligible range. As seen in Table 46, the difficulty classifications (Easy, Average, or Difficult) were the same for children with and without an IEP for seven of the nine progressions.

Physical

The Rasch-Welch *t* was statistically significant for all five progressions within the physical domain. The Mantel-Haenszel χ^2 statistic was also statistically significant for all five progressions. However, the magnitude of the DIF contrasts was negligible for all nine progressions. The differences in estimated difficulties were very small in absolute value (.09 to .23) for all five progressions. As seen in Table 47, the difficulty classifications (Easy, Average, or Difficult) were the same for children with and without an IEP for three of the five progressions.

Language

The Rasch-Welch *t* was statistically significant for all eight progressions within the language domain. The Mantel-Haenszel v statistic was also statistically significant for all eight progressions. However, the magnitude of the DIF contrasts were negligible for seven of the eight progressions. The DIF contrast was intermediate (.48) in magnitude for one progression, Objective 8a. The differences in estimated difficulties were very small in absolute value (.07 to .28) for five of eight progressions. Only two progressions, Objectives 8b and 9d, yielded a DIF contrast statistic (.37 and .44) that was close to the upper end of the negligible range. As seen in Table 48, the difficulty classifications (Easy, Average, or Difficult) were the same for children with and without an IEP for seven of the nine progressions.

Cognitive

The Rasch-Welch *t* was statistically significant for 8 of the 10 progressions within the cognitive domain. The Mantel-Haenszel χ^2 statistic was statistically significant for 7 of the 10 progressions. However, the magnitude of the DIF contrasts were negligible for all 10 progressions. The differences in estimated difficulties were very small in absolute value (.00 to .25) for all 10 progressions. As seen in Table 49, the difficulty classifications (Easy, Average, or Difficult) were the same for children with and without an IEP for 9 of the 10 progressions.

Literacy

The Rasch-Welch *t* was statistically significant for 15 of the 16 progressions within the literacy domain. The Mantel-Haenszel χ^2 statistic was statistically significant for 14 of the 16 progressions. However, the magnitude of the DIF contrasts was negligible for all 16 progressions. The differences in estimated difficulties were very small in absolute value (.02 to .34) for 14 of the 16 progressions. Only two progressions, Objectives 16a and 18e, yielded DIF contrast statistics (.37 and .42)

that were close to the upper end of the negligible range. As seen in Table 50, the difficulty classifications (Easy, Average, or Difficult) were the same for children with and without an IEP for 15 of the 16 progressions.

Mathematics

The Rasch-Welch *t* was statistically significant for all 12 progressions within the mathematics domain. The Mantel-Haenszel χ^2 statistic was statistically significant for 10 of the 12 progressions. However, the magnitude of the DIF contrasts was negligible for all 12 progressions. The differences in estimated difficulties were very small in absolute value (.02 to .35) for all 12 progressions. As seen in Table 51, the difficulty classifications (Easy, Average, or Difficult) were the same for children with and without an IEP for 11 of the 12 progressions.

Summary of Current DIF Analyses

The estimated average difficulties for female and male children were identical for many *GOLD*® progressions. Where small differences were found, the magnitude of those differences was considered negligible for all of the progressions. The difficulty categories (Easy, Average, or Difficult) were the same for female and male children across all but one of the *GOLD*® progressions. Therefore, the results of this study yielded no substantive evidence to support a conclusion that DIF based on gender is present as teachers use the *GOLD*® developmental progressions.

The estimated average difficulties for native English (NE) speakers and English- language learners (ELLs) were identical for most *GOLD*® progressions. Where small differences were found, the magnitude of those differences was considered negligible for all of the progressions. The difficulty categories (Easy, Average, or Difficult) were the same for NE and ELL children for all but one of the *GOLD*® progressions. Therefore, the results of this study yielded no substantive evidence to support a conclusion that DIF based on primary language is problematic as teachers use the *GOLD*® developmental progressions.

The estimated average difficulties for white and non-white children were identical for many *GOLD*® progressions. Where small differences were found, the magnitude of those differences was considered negligible for all of the progressions. The difficulty categories (Easy, Average, or Difficult) for white and non-white children were the same across all but one of the *GOLD*® progressions. Therefore, the results of this study yielded no substantive evidence to support a conclusion that DIF based on race/ethnicity is present as teachers use the *GOLD*® developmental progressions. However, for two of the progressions, the differences in estimated difficulty were at the upper end of the negligible range. These two progressions, Objectives 18d (DIF contrast = .39) and 19c (DIF contrast = .41), will need to be monitored carefully in future research.

It may be useful to review the content of the behavioral anchors associated with each step on those two progressions, Objectives 18d and 19c, as some differences were found between the estimated difficulties for white and non-white children. The content of these progressions could be reviewed for fairness to all racial/ethnic subgroups of children. It is also possible that teachers could benefit from more training, specific to Objectives 18d and 19c, on how to recognize, elicit, and analyze evidence of child developmental progress from non-white children.

With respect to IEP status, this study contributes validity evidence that supports the appropriate application of *GOLD*® as a formative assessment measure for students with disabilities. These findings suggest there is almost no evidence of DIF in five of the six domains included in this study, with 59 of the 60 progressions demonstrating negligible DIF. The only exception is Objective 8a. Objective 8a focuses on comprehending increasingly complex oral language, and this progression was easier for children with an IEP. These findings are different from previous studies exploring DIF analysis for this subgroup (Kim et al., 2013; Lambert et al., 2014). In previous research, a few progressions related

to a child's ability to use language to express thoughts and needs yielded higher difficulty levels for children with an IEP. In contrast, findings from this study suggest teachers may be lenient toward children with disabilities when making placements on this progression or have decreased expectations for their receptive communication skills.

Alternatively, given the experience of educators working with children with disabilities in inclusive Head Start and state-funded pre-K programs, teachers may be more sensitive to the language needs and functioning of children with disabilities. For example, North Carolina requires Birth to Kindergarten licensure for pre-K teachers, including special education training. This suggests these teachers may have additional skill sets for effective communication with students with diverse needs.

Findings from this study can contribute to stakeholder confidence in the appropriate use of *GOLD*® for students with disabilities. In addition to contributing to validity expectations in demonstrating fairness of the measure as outlined in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), potentially the most significant benefit of increasing stakeholder confidence is increased utility of assessment outcomes for enhanced instructional planning, resulting in improved outcomes for students with disabilities.

It is important to note the potential limitations of this research. The analyses that addressed primary language used one specific ELL subgroup as the focal group and NE children as the reference group. Therefore, the results of this study are limited to those specific groups. Future studies could attempt to examine potential DIF with other linguistic and cultural subgroups of children. In addition, this study treated all ELL children as a unitary group without respect to levels of acculturation or language acquisition. Future research could benefit from a more nuanced examination of subgroups within the ELL population (Kim, Lambert, Durhams, & Burt, 2018).

Similarly, the analyses that addressed race and ethnicity were limited to non-white children as the focal group and white children as the reference group. Future analyses could attempt to examine whether the results of this study extend to additional subgroups based on race or ethnicity. Furthermore, this study did not incorporate any information about the gender, native language, or race/ethnicity of the teachers that made the placements on *GOLD*® progressions. Future research could include demographic information about the teachers, along with racial and ethnic congruence between teachers and the children and families they serve.

Limitations to the examination of DIF by IEP status include the composition of teachers in the sample. The sample comprises children taught by teachers within programs that serve a higher concentration of children with special needs than typical prekindergarten programs. In addition, state-funded pre-K and Head Start classrooms are designed to meet the unique needs of students with special needs. Therefore, these results may not be generalized to all teachers. Future research with a more nationally representative sample is recommended.

When the results generated by this research are taken as a whole, they offer very little evidence that the *GOLD*® progressions are measuring different latent constructs for the focal and reference groups compared. These results generally confirm the findings of previous studies (Kim et al., 2013; Kim et al., 2014; Kim et al., 2018). Therefore, it is reasonable to conclude that teachers are generally using the *GOLD*® progressions to make ratings of children in a similar manner across the subgroups compared in this study. At least with respect to the subgroups compared, the evidence of these DIF analyses suggest it is reasonable to interpret the resulting assessment scores in a similar manner for these subgroups of children. However, it is important to note that teachers can introduce other sources of construct irrelevant variance and rater effects when engaging in the complex response process involved with authentic formative assessment for young children. Therefore, future research can focus on furthering the validity argument for the use of *GOLD*® with young children, including examination of evidence for inter-rater reliability.

Establishing Stakeholder Confidence

Establishing stakeholder confidence that an educational assessment is appropriate for use with a specific population or subgroup, and for a well-defined purpose, is necessary to ensure educators will use the information a measure provides for making appropriate instructional decisions. When educators question the value of an assessment, they are less likely to apply the information a measure provides to inform their instructional practice (Ackerman, 2018; Harvey & Ohle, 2018, Holcomb et al., 2020). Given the individualized learning needs of students with developmental disabilities, it is natural for educators to question whether a commercially available assessment can yield valuable information that captures these children's unique skills and abilities.

The *Standards for Educational and Psychological Testing* posit that gathering validity evidence is the responsibility of the test developer and those interpreting test outcomes (2014). For an authentic formative assessment like *GOLD*®, teachers are essential to ensuring the validity of scores, including their proper use and application within the instructional process. Therefore, examination of the response process is an important aspect of the validity evidence gathering process.

Authentic formative assessment requires the teacher to play a central role in the response process. The teacher makes critical decisions about where to place students on the developmental progressions based on their interpretation of selected evidence. For an assessment to be reliable and fair, teachers must consistently apply each developmental rating scale across all subgroups, including children with disabilities. The Standards indicate "validation may include empirical studies of how observers or judges record and evaluate data along with analyses of the appropriateness of these processes to the intended interpretation or construct definition" (AERA, APA, & NCME, 2014, p.16).

Prevalence of *GOLD*® in Assessing Children With Special Needs

States and local school districts are required by law to provide inclusive preschool experiences for children between the ages of three and five years old who have been identified with one or more qualifying disabilities. In support of this legislation, federally funded Head Start and Early Head Start programs stipulate 10% of the total funded enrollment must be children with special needs, which resulted in services for more than 260,000 children with disabilities in 2017-2018 (ECLKC, 2020). Head Start programs deliver services through 1,600 agencies in local communities and provide services to more than one million children every year (Office of Head Start, 2020). In addition, state and locally funded prekindergarten programs in many states also serve children with disabilities.

The Head Start Child Outcomes Framework: Ages Birth to Five (ELOF) provides guidance for developing and sustaining pre-K programs and instructional practices that promote successful learning. Designed to include children from diverse linguistic, economic, and cultural backgrounds and children with disabilities, the integrated framework addresses five broad areas of learning that are organized into domains, sub-domains, goals, developmental progressions, and indicators. For young children with special needs, the framework fosters collaboration between families, teachers, and specialists identified on the child's Individual Family Service Plans (IFSP) or Individual Educational Plan (IEP).

Ensuring delivery of high-quality pre-K experiences for children requires a coordinated and integrated system of assessments, including screening measures, formative and summative child assessments, measures of environmental quality, and measures of adult-child interactions. The Individuals with Disabilities Education Act (IDEA) requires states to annually report outcomes for pre-K children with disabilities to the Office of Special Education Planning (OSEP). This comprehensive assessment system requires the use of ongoing assessment measures designed to inform instructional practices and evaluate children's progress in key areas of development (Atkins-Burnett et al., 2014).

Observation-based authentic formative assessment is widely used across the United States to assess the developmental status of young children, including the evaluation of kindergarten

readiness. Research indicates that using authentic formative assessment (AFA) can lead to significant improvement in classroom quality (Hattie & Timperley, 2007; Stover et al., 2013; Torrence, 2001). For this reason, the National Association for the Education of Young Children and the Council for Exceptional Children's Division for Early Childhood endorse authentic formative assessment approaches. AFA measures have grown in popularity in early childhood education programs largely because they provide unique advantages over traditional standardized assessment practices. Unlike standardized assessments that are summative in nature and often administered under artificial contexts and protocols, AFA has the potential to transform educators' practice in real time, resulting in improved educational opportunities for students. Observations in naturalistic settings allow children to demonstrate skills and behaviors in a variety of ways, across a variety of contexts, providing educators with a more comprehensive picture (Neisworth & Bagnato, 2004). Information rendered from AFA measures facilitates next steps in instructional practices, strengthens rich communication and collaboration with families and other professionals, and provides unique coaching opportunities for targeted professional development to enhance teacher practices. Ultimately, AFA provides greater capacity for usefulness to educators than traditional standardized assessments (Lambert, 2020).

Validity Evidence Based on Internal Structure

Confirmatory Factor Analysis

Validity evidence based on the internal structure of the item-level information a measure provides is one of the most important components of an argument to support the meaningfulness, usefulness, and appropriateness of the interpretations of assessment scores (Messick, 1995; AERA, APA, & NCME, 2014). Standard 1.13 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014, p. 26) states, "If the rationale for a test score interpretation for a given use depends on premises about the relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided." Given that the information provided by *GOLD*® progressions can be organized and reported by domain of development in the form of scaled scores, this research phase focused on gathering additional validity evidence for the use of *GOLD*® based on relationships between the progressions.

Previous research presented this type of validity evidence for the birth to kindergarten edition of *GOLD*® (Lambert, Kim, & Burts, 2015). We used confirmatory factor analysis (CFA), a special case of Structural equation modeling (SEM), to replicate those findings with the birth to third grade version of *GOLD*® and to provide evidence to support the hypothesized factor structure. CFA can provide supporting evidence for the organization of observed scores, the placements on *GOLD*® progressions, into domain-specific factor scores. CFA allows researchers to examine magnitude and direction of the relationships between observed variables and their underlying latent constructs with a particular focus on whether sample data fits the proposed factor model. We used the Lavaan package within the R software program (Rosseel, 2012) to test the hypothesized domain-specific measurement models and overall factor models.

We assessed model fit with the Comparative Fit Index (CFI), Tucker Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR). Several authors have offered useful guidelines for judging adequacy of model fit by evaluating fit indexes. For example, several have suggested fit indexes of .90 or higher indicate adequate fit (Bentler & Bonett, 1980; Bollen, 1989). Hooper, Coughlan, and Mullen (2008) suggested that values of .95 for the CFI and TFI indicate good fit. For the SRMR, cut-off values of .06 (Hu & Bentler, 1999) or .07 (Steiger, 2007) have been suggested as indicators of good fit. The RMSEA has the added advantage of providing a confidence interval around its value (Steiger, 1990; MacCallum, Browne, & Sugawara, 1996), which allows for the null hypothesis of inadequate fit to be tested. It has been suggested that, in a model with good fit, the RMSEA lower limit should be close to 0 and the upper limit less than 0.08 (McQuitty, 2004). Hu and Bentler (1999) suggest a two-stage assessment of fit that requires CFI or TFI values of .96 or greater and SRMR values less than .09 to support an interpretation of good fit.

The first step in assessing the fit of the data to the proposed factor structure was fitting measurement models for each of the six *GOLD*® domains of development. These models test whether the progressions within a given domain fit together to represent a single underlying latent construct. This process involved testing a simple model for each domain that contained one latent variable, which represents the true underlying level of developmental progress for the domain under study (social–emotional, physical, language, cognitive, literacy, or mathematics), and pathways from the latent variable to each of the observed variables. The observed variables in these models are the placements on the *GOLD*® progressions for each of the progressions within the respective domain. The *GOLD*® progressions are comprised of ordinal rating scales. The numbers assigned to the placements made by teachers signify ordered levels, not exact quantities. Therefore, all parameters were estimated using the Diagonally Weighted Least Squares estimation procedure within the Lavaan package.

The results of testing the measurement model for the social–emotional domain confirmed the unidimensionality of the respective scaled score. The fit statistics were strong: CFI = .986, TLI = .980, RMSEA = .095 (90% CI = .093 to .097), and SRMR = .009. The RMSEA value was slightly above the acceptable range, but the other indicators of fit were well within acceptable limits. Each of the pathways from the latent variable to the *GOLD*® progressions was statistically significant. These path coefficients were all high (.892 - .956). Two correlated error terms emerged. The residual variances for Objectives 1c and 2a were correlated (.378). These progressions address a child's ability to meet their own needs and form relationships with adults, suggesting potential unmeasured factors related to relying on adults. The residual variances for Objectives 2c and 2d were correlated (.287). These progressions address a child's ability to form relationships with other children.

The results of testing the measurement model for the physical domain confirmed the unidimensionality of the respective scaled score. The fit statistics were strong: CFI = .999, TLI = .998, RMSEA = .038 (90% CI = .033 to .043), and SRMR = .002. Therefore, all indicators of fit were well within acceptable limits. Each of the pathways from the latent variable to the *GOLD*® progressions was statistically significant. These path coefficients were all high (.992 - .963). One correlated error term emerged. The residual variances for Objectives 7a and 7b were correlated (.314). These progressions address a child's fine-motor skills.

The results of testing the measurement model for the language domain confirmed the unidimensionality of the respective scaled score. The fit statistics were strong: CFI = .984, TLI = .978, RMSEA = .111 (90% CI = .107 to .112), and SRMR = .008. The RMSEA value was above the acceptable range, but the other indicators of fit were well within acceptable limits. Each of the pathways from the latent variable to the *GOLD*® progressions was statistically significant. These path coefficients were all high (.945 - .966).

The results of testing the measurement model for the cognitive domain confirmed the unidimensionality of the respective scaled score. The fit statistics were strong: CFI = .985, TLI = .980, RMSEA = .095 (90% CI = .093 to .097), and SRMR = .007. The RMSEA value was slightly above the acceptable range, but the other indicators of fit were well within acceptable limits. Each of the pathways from the latent variable to the *GOLD*® progressions was statistically significant. These path coefficients were all high (.945 - .966). One correlated error term emerged. The residual variances for Objectives 14a and 14b were correlated (.245). These progressions address a child's ability to think symbolically and engage in sociodramatic play.

The results of testing the measurement model for the literacy domain confirmed the unidimensionality of the respective scaled score. The fit statistics were strong: CFI = .999, TLI = .999, RMSEA = .078 (90% CI = .077 to .079), and SRMR = .054. Therefore, all indicators of fit were well within acceptable limits. Each of the pathways from the latent variable to the *GOLD*® progressions

was statistically significant. These path coefficients were all high (.885 - .980). One correlated error term emerged. The residual variances for Objectives 18d and 18e were correlated (.757). These progressions address a child's ability to read and comprehend text.

The results of testing the measurement model for the mathematics domain confirmed the unidimensionality of the respective scaled score. The fit statistics were strong: CFI = .999, TLI = .999, RMSEA = .029 (90% CI = .027 to .030), and SRMR = .029. Therefore, all indicators of fit were well within acceptable limits. Each of the pathways from the latent variable to the *GOLD*® progressions was statistically significant. These path coefficients were all high (.821 - .991). Three correlated error terms emerged. The residual variances for Objectives 20d, 20e, and 20f were correlated (.851, .835, .890). These progressions address a child's ability to apply mathematical operations.

Next, the measurement models tested simultaneously within an overall factor model. The fit statistics were strong: CFI = .997, TLI = .997, RMSEA = .051 (90% CI = .051 to .051), and SRMR = .043. Therefore, all indicators of fit were well within acceptable limits. Each of the pathways from the latent variables to the *GOLD*® progressions remained statistically significant. These path coefficients were all high (.816 - .987) and similar in magnitude to their corresponding estimates from the measurement models. All of the pathways between the six latent variables were high (> .880), suggesting strong correlations between the latent variables. Therefore, we tested a second factor model that contained a higher factor we called "Child Developmental Progress." This model tested whether an overall latent construct that represents the overall developmental level of the child might account for the strong relationships between the latent variables. The fit statistics were strong: CFI = .997, TLI = .997, RMSEA = .055 (90% CI = .054 to .055), and SRMR = .048. Therefore, all indicators of fit were well within acceptable limits. Each of the pathways from the latent variables to the *GOLD*® progressions remained statistically significant. These path coefficients were all high (.810 - .987) and similar in magnitude to their corresponding estimates from the first factor model. All of the pathways between the higher order factor and the six latent variables were high (> .900), suggesting the strong correlations between the latent variables could be explained by the higher order factor.

Measurement Invariance

Standard 3.11 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014, p. 63) states "Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the population." Extensive examination of DIF by subgroups was already summarized earlier in this manual. To supplement and expand upon those findings, we conducted a series of measurement invariance analyses using the same subgroups. This phase of the research focused on gathering additional validity evidence to support the premise that *GOLD*® scaled scores can be interpreted and used in the same way across various subgroups of children.

Researchers can use measurement invariance analyses to evaluate the psychometric equivalence of assessment scores from the same measure across subgroups. Evidence to support measurement invariance implies that the scores in question have the same meaning across the subgroups. Conversely, non-invariance suggests the information an assessment provides may not have the same internal structure across subgroups, and therefore scaled scores may not lead to the same interpretations across subgroups. In a broad sense, the research question at stake in measurement invariance analyses focuses on whether observed group differences are associated with actual differences between subgroups or with differences in how the assessment measure functions across the subgroups.

Evidence of measurement invariance, therefore, means researchers can make claims about specific types of statistical equivalence among scores from various subgroups from the intended

population. Four types of statistical equivalence are tested within a multi-stage measurement invariance analytical process (Beaujean, 2014). First, researchers examine Configural Invariance. Researchers test the hypothesized factor model separately for each subgroup. This stage can provide evidence that the hypothesized factor structure is supported for the subgroups in question. Second, researchers examine Weak Invariance. Researchers introduce constraints into the model to fix the factor loading to be equal across the subgroups. This stage can provide evidence that the factor loadings are equivalent across the subgroups. Next, researchers examine Strong Invariance. Researchers introduce constraints into the model to fix the factor loading and intercepts to be equal across the subgroups. This stage can provide evidence that the intercepts are equivalent across the subgroups. Finally, researchers examine Strict Invariance. Researchers introduce constraints into the model to fix the factor loading, intercepts, and error variances to be equal across the subgroups. This stage can provide evidence that the error variances are equivalent across subgroups.

We conducted all measurement invariance analyses using the Lavaan package within R. We assessed how well the data fit the nested hypothesized models using the same goodness of fit indexes applied to the CFA analyses. Evidence of fit between the data and the hypothesized models is evidence to support measurement invariance across the subgroups in question. In addition, if fit indexes decline as additional constraints are imposed within the models, additional testing of the magnitude of the changes between nested models can be applied (Widaman & Thompson, 2003).

With and Without an IEP

We examined measurement invariance between children with and without an IEP. We used the same dataset described within the section of this manual regarding DIF analyses by disability status subgroups. As shown in Table 53, the fit statistics for the test of Configural Invariance were strong for children without an IEP: CFI = .988, TLI = .987, RMSEA = .067 (90% CI = .067 to .067), and SRMR = .063. The fit statistics were also strong for children with an IEP: CFI = .981, TLI = .980, RMSEA = .079 (90% CI = .079 to .080), and SRMR = .076. Therefore, all indicators of fit were within acceptable limits for both subgroups, providing support for the hypothesized factor structure across both subgroups.

The fit statistics for the test of Weak Invariance were strong: CFI = .985, TLI = .985, RMSEA = .074 (90% CI = .074 to .074), and SRMR = .071. Therefore, all indicators of fit were within acceptable limits for both subgroups, providing support for the equivalence of factor loadings across the subgroups. The fit statistics for the test of Strong Invariance were also strong: CFI = .984, TLI = .985, RMSEA = .073 (90% CI = .073 to .074), and SRMR = .070. Therefore, all indicators of fit were within acceptable limits for both subgroups, providing support for the equivalence of factor loadings and intercepts across the subgroups. The fit statistics for the test of Strict Invariance were also strong: CFI = .984, TLI = .985, RMSEA = .073 (90% CI = .073 to .073), and SRMR = .070. Therefore, all indicators of fit were within acceptable limits for both subgroups, providing support for the equivalence of factor loadings, intercepts, and residual variances across the subgroups.

Gender

We examined measurement invariance between male and female children. We used the same dataset described within the section of this manual regarding DIF analyses by gender. As shown in Table 54, the fit statistics for the test of Configural Invariance were strong for male children: CFI = .997, TLI = .997, RMSEA = .053 (90% CI = .053 to .053), and SRMR = .045. The fit statistics were also strong for female children: CFI = .998, TLI = .997, RMSEA = .048 (90% CI = .048 to .048), and SRMR = .042. Therefore, all indicators of fit were within acceptable limits for both subgroups, providing support for the hypothesized factor structure across both subgroups.

The fit statistics for the test of Weak Invariance were strong: CFI = .997, TLI = .997, RMSEA = .051 (90% CI = .051 to .051), and SRMR = .045. Therefore, all indicators of fit were within acceptable limits

for both subgroups, providing support for the equivalence of factor loadings across the subgroups. The fit statistics for the test of Strong Invariance were also strong: CFI = .997, TLI = .997, RMSEA = .050 (90% CI = .049 to .050), and SRMR = .044. Therefore, all indicators of fit were within acceptable limits for both subgroups, providing support for the equivalence of factor loadings and intercepts across the subgroups. The fit statistics for the test of Strict Invariance were also strong: CFI = .997, TLI = .997, RMSEA = .049 (90% CI = .049 to .049), and SRMR = .044. Therefore, all indicators of fit were within acceptable limits for both subgroups, providing support for the equivalence of factor loadings, intercepts, and residual variances across the subgroups.

Language Spoken in the Home

We examined measurement invariance between children from homes where English is spoken and where Spanish is spoken. We used the same dataset described within the section of this manual regarding DIF analyses by language spoken in the home. As shown in Table 55, the fit statistics for the test of Configural Invariance were strong for children from English speaking homes: CFI = .997, TLI = .997, RMSEA = .051 (90% CI = .051 to .051), and SRMR = .043. The fit statistics were also strong for children from Spanish speaking homes: CFI = .992, TLI = .992, RMSEA = .062 (90% CI = .061 to .063), and SRMR = .060. Therefore, all indicators of fit were within acceptable limits for both subgroups, providing support for the hypothesized factor structure across both subgroups.

The fit statistics for the test of Weak Invariance were strong: CFI = .993, TLI = .993, RMSEA = .077 (90% CI = .077 to .077), and SRMR = .056. Therefore, all indicators of fit were within acceptable limits for both subgroups, providing support for the equivalence of factor loadings across the subgroups. The fit statistics for the test of Strong Invariance were also strong: CFI = .996, TLI = .996, RMSEA = .057 (90% CI = .057 to .057), and SRMR = .049. Therefore, all indicators of fit were within acceptable limits for both subgroups, providing support for the equivalence of factor loadings and intercepts across the subgroups. The fit statistics for the test of Strict Invariance were also strong: CFI = .996, TLI = .996, RMSEA = .057 (90% CI = .057 to .057), and SRMR = .050. Therefore, all indicators of fit were within acceptable limits for both subgroups, providing support for the equivalence of factor loadings, intercepts, and residual variances across the subgroups.

Race/Ethnicity

We examined measurement invariance between white children and children of color (Hispanic or African American). We used the same dataset described within the section of this manual regarding DIF analyses by race/ethnicity. As shown in Table 56, the fit statistics for the test of Configural Invariance were strong for white children: CFI = .998, TLI = .998, RMSEA = .045 (90% CI = .045 to .045), and SRMR = .038. The fit statistics were also strong for children of color: CFI = .995, TLI = .995, RMSEA = .054 (90% CI = .054 to .055), and SRMR = .049. Therefore, all indicators of fit were within acceptable limits for both subgroups, providing support for the hypothesized factor structure across both subgroups.

The fit statistics for the test of Weak Invariance were strong: CFI = .994, TLI = .994, RMSEA = .074 (90% CI = .073 to .074), and SRMR = .069. Therefore, all indicators of fit were within acceptable limits for both subgroups, providing support for the equivalence of factor loadings across the subgroups. The fit statistics for the test of Strong Invariance were also strong: CFI = .997, TLI = .997, RMSEA = .053 (90% CI = .053 to .054), and SRMR = .050. Therefore, all indicators of fit were within acceptable limits for both subgroups, providing support for the equivalence of factor loadings and intercepts across the subgroups. The fit statistics for the test of Strict Invariance were also strong: CFI = .997, TLI = .997, RMSEA = .053 (90% CI = .053 to .054), and SRMR = .051. Therefore, all indicators of fit were within acceptable limits for both subgroups, providing support for the equivalence of factor loadings, intercepts, and residual variances across the subgroups.

Summary of Measurement Invariance Analyses

The hypothesized factor structure was sustained for each of the measurement invariance

analyses, across all subgroups examined. The measurement invariance analyses demonstrated acceptable fit for each stage of process for all subgroups. There were no stages within any of the subgroup analyses that yielded a substantial reduction in fit relative to the results of a previous stage. Therefore, no further testing is necessary. These results offer strong evidence that the *GOLD*® assessment scores can be interpreted in the same manner across all subgroups analyzed.

Overall Summary

The primary goal of this study was to extend the reliability and validity evidence for *GOLD*® using a nationally representative sample of children served by teachers who assess children in fall, winter, and spring of the academic year. Overall, this study reaffirmed that the use of *GOLD*® with young children continues to yield highly reliable scaled scores as indicated by both the classical and Rasch reliability statistics. The results also demonstrate strong statistical evidence that the developmental progressions within each scale generally work very well together to measure a single underlying domain of development. The results further demonstrate evidence that the ratings can be successfully organized within each developmental domain generally as intended by the team that originally developed the assessment. There is also statistical evidence that teachers are able to use the rating scale effectively to place children along a progression of development and learning. When the progressions within each domain are arranged from the easiest to the most difficult objectives for children to master, the hierarchy that is created generally matches very well with what developmental theory indicates. Therefore, the range of difficulties indicates that each section of *GOLD*® can be used by teachers to help them understand the developmental trajectory that most children will follow.

The CFA analyses demonstrated strong evidence to support the intended *GOLD*® factor structure. Furthermore, the CFA results suggest the presence of an overall, or higher order factor, that contributes to the *GOLD*® scaled scores. This manual also contains extensive analyses of subgroup differences. Across all subgroups considered, reliability coefficients remained very consistent. DIF analyses demonstrated very little evidence for potential biases or differences in difficulty across all subgroups analyzed. In addition, measurement invariance analyses demonstrated strong evidence that the *GOLD*® scaled scores can be used and interpreted in the same way across all subgroups considered.

Data from a wider range of second- and third-grade children is needed for future studies, especially given the relatively small numbers of children in these grade levels in the study, and the relatively smaller numbers of children placed at the upper ends of many of the progressions. Future research could focus on measures of the degree of association between *GOLD*® scaled scores and external measures of child developmental progress. It would also be helpful to conduct a study that addresses inter-rater reliability. Follow-up studies should focus on both procedural fidelity and agreement with expert raters as well as variance decomposition methods that address generalizability. As teachers around the country gain more experience and training with the use of the assessment, it may also be helpful to conduct studies that examine the proportion of the variability in ratings that is between and within raters, the sensitivity of the scores to growth over time, and differences between subgroups of children. Lastly, future research is needed to evaluate whether teachers are collecting sufficient quantities of high-quality, valid evidence of child growth, development, and learning to support placements on the *GOLD*® developmental progressions.

Teachers are being asked to implement a variety of assessments and assessment-related tasks. They must do so while facing challenges related to limited instructional resources, many demands on their time, and increasingly diverse classrooms of children with individualized needs. All of this comes at a time when teachers are often unprepared for the linguistic and cultural diversity that is the reality in most classrooms. To plan instruction accordingly, they need assessments to help them document and understand the developmental status, strengths, needs, and growth patterns of children. This study has robustly demonstrated that *GOLD*® satisfies the requirement for high-quality assessment in early childhood education programs.

References

Ackerman, D. J. (2018). Real world compromises: Policy and practice impacts of kindergarten entry assessment-related validity and reliability challenges. Research Report. ETS RR-18-13. *ETS Research Report Series*.

Andrich, D. (1978). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, 2, 581–594.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-574.

Atkins-Burnett, S., Monahan, S., Akers, L., Carta, J., Wasik, B. A., & Boller, K. (2014). Conceptual framework and measurement plan. *Tailored eaching: Teachers' use of ongoing child assessment to individualize instruction* (Vol. 1). U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation.

Badia, X., Prieto, L., and Linacre, J. M. (2002). Differential item and test functioning (DIF and DTF). *Rasch Measurement Transactions*, 16, 889.

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. Routledge.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum Associates Publishers.

Early Childhood Learning & Knowledge Center. (2020, December). *The office of Head Start program information report*. https://eclkc.ohs.acf.hhs.gov/data-ongoing-monitoring/article/program-information-report-pir

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for all students: Development, implementation, and analysis* (pp. 261-287). Routledge. https://doi.org/10.4324/9781410605115

Engelhard, G., & Wind, S. (2017). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.

Gallagher, P. A., & Lambert, R. G. (2006). Classroom quality, concentration of children with special needs, and child outcomes in Head Start. *Exceptional Children*, 73(1), 31-52.

Harvey, H., & Ohle, K. (2018). What's the purpose? Educators' perceptions and use of a state-mandated kindergarten entry assessment. *Education Policy Analysis Archives*, 26, 41.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112.

Heritage, M. (2013). *Formative assessment in practice: A process of inquiry and action*. Boston: Harvard Education Press.

Heroman, C., Burts, D.C., Berke, K., & Bickart, T.S. (2010). *Teaching Strategies GOLD® objectives for development & learning: Birth through Kindergarten*. Bethesda, MD: Teaching Strategies, LLC

Holcomb, T. S., Li, Z., Ferrara, A. M., & Lambert, R. G. (2020). *Practitioner perspectives on the implementation of the North Carolina kindergarten entry assessment*. (CEMETR-2020-11). Center for Educational Measurement and Evaluation.

Joint Committee on the *Standards for Educational and Psychological Testing* of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (2014). *The Standards for Educational and Psychological Testing*. American Educational Research Association.

Joseph, G., Soderberg, J.S., Stull, S., Cummings, K., McCutchen, D. & Han, R.J. (2020). Inter-rater reliability of Washington State's kindergarten entry assessment. *Early Education and Development*, 31(5), 764-777. DOI: 10.1080/10409289.2019.1674589.

Kim, D-H., Lambert, R.G., & Burts, D.C. (2013). Evidence of the validity of *Teaching Strategies GOLD®* assessment tool for English language learners and children with disabilities. *Early Education and Development*, 24(4), 574–595. https://doi.org/10.1080/10409289.2012.701500

Kim, D-H., Lambert, R. G., & Burts, D. C. (2014). Validating a developmental scale for young children using the Rasch Model: Applicability of the *Teaching Strategies GOLD®* assessment system. *Journal of Applied Measurement*, 15(4), 405–421.

Kim, D-H., Lambert, R.G., Durham, S., & Burts, D. (2018). Examining the validity of *GOLD®* with 4-year-old dual language learners. *Early Education and Development*, 29(4), 477–493. https://doi.org/10.1080/10409289.2018.1460125

Lambert, M. C., Garcia, A. G., January, S-A.A., & Epstein, M. H. (2017). The impact of English language learner status on screening for emotional and behavioral disorders: A differential item functioning (DIF) study. *Psychology in the Schools*, 55(3), 229-239. https://doi.org/10.1002/pits.22103

Lambert, R. (2020). *Technical manual for the Teaching Strategies GOLD® assessment (2nd edition): Birth through third grade*. Center for Educational Measurement and Evaluation, University of North Carolina at Charlotte.

Lambert, R. (2020). Shaping a validity argument for the use of authentic formative assessments to support young children. In C. Martin, D. Polly, & R. Lambert (Eds.), *Handbook of research on formative assessment in pre-K through elementary Classrooms* (pp. 49-69). IGI-Global.

Lambert, R. (2017). *Technical manual for the* Teaching Strategies GOLD™ *assessment system: Birth through third grade edition* (CEMETR-2017-02). Center for Educational Measurement and Evaluation.

Lambert, R., Kim, D-H., & Burts, D. (2015). The measurement properties of the *Teaching Strategies GOLD® assessment system. Early Childhood Research Quarterly*, 33(4), 49-63.

Lambert, R. G., Kim, D-H., & Burts, D. C. (2014). Using teacher ratings to track the growth and development of young children using the *Teaching Strategies GOLD®* assessment system. *Journal of Psychoeducational Assessment*, 32(1), 27–39.

Lambert, R., Kim, D-H., & Burts, D. (2014). *Technical manual (3rd edition) for the* Teaching Strategies *GOLD® assessment system* (CEMETR-2014-06). Center for Educational Measurement and Evaluation.

Lambert, R., Kim, D-H., & Burts, D. (2013). *Technical manual (2nd edition) for the* Teaching Strategies *GOLD® assessment system* (CEMETR-2013-04). Center for Educational Measurement and Evaluation.

Lambert, R. & Kim, D-H., Taylor, H., & McGee, J. (2010). *Technical manual for the* Teaching Strategies GOLD™ *assessment system* (CEMETR-2010-06). Center for Educational Measurement and Evaluation.

Lambert, R., Kim, D-H., & Burts, D. (2013). *Evidence for the association between scores from the Teaching Strategies* GOLD® *assessment system and information from direct assessments of child progress*. (CEMETR-2013-05). Center for Educational Measurement and Evaluation.

Lambert, R. G., Kim, D-H., & Burts, D. C. (2013). Using teacher ratings to track the growth and development of young children using the *Teaching Strategies GOLD®* assessment system. *Journal of Psychoeducational Assessment*, 32(1), 27-39. https://doi.org/10.1177/0734282913485214

Linacre, J. M. (2012). Winsteps (Version 3.75.1) [Computer Software]. Chicago, IL: Winsteps.com.

Linacre, J.M. (2020). Winsteps (Version 4.6.2) [Computer software]. Beaverton, OR: Winsteps.com.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psycholo*gist, 50(9), 741-749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Miller-Bains, K. L., Russo, J. M., Williford, A. P., DeCoster, J., & Cottone, E. A. (2017). Examining the validity of a multidimensional performance-based assessment at kindergarten entry. *AERA Open*, 3(2), 1–16.

Mitchell, S. (2016). *Comprehensive assessment systems in preschool development and expansion grant states. State technical assistance report*. Preschool Development and Expansion Grant Technical Assistance.

Neisworth, J. T., & Bagnato, S. J. (2004). The mismeasure of young children: The authentic assessment alternative. *Infants & Young Children*, 17(3), 198-212.

Office of Head Start. (2019). National Head Start services snapshot. Retrieved from https://eclkc.ohs.acf.hhs.gov/sites/default/files/pdf/no-search/service-snapshot-hs-2018-2019.pdf

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (3rd ed.). SAGE Publications.

Qi, C.H., Marley, S.C. (2009). Differential item functioning analysis of the preschool language scale-4 between English-speaking Hispanic and European American children from low-income families. *Topics in Early Childhood Special Education*, 29(3), 171-180. https://doi.org/10.1177/0271121409332674

Russo, J. M., Williford, A. P., Markowitz, A. J., Vitiello, V. E., & Bassok, D. (2019). Examining the validity of a widely-used school readiness assessment: Implications for teachers and early childhood programs. *Early Childhood Research Quarterly*, 48(3), 14–25. https://doi.org/10.1016/j.ecresq.2019.02.003

Stover, K., Wood, K. D., Donovan, E., Paratore, J. R., & McCormack, R. (2013). A tale of two authentic assessment tools. *Voices from the Middle*, 21(2), 26-35. https://doi.org/10.58680/vm201324463

Torrance, H. (2001) . Assessment for learning: Developing formative assessment in the classroom. *Education 3-13*, 29(3), 26–32 .

U.S. Department of Health and Human Services, Administration for Children and Families, Office of Head Start. (2010). *The Head Start child development and early learning framework: Promoting positive outcomes in early childhood programs serving children 3–5 years old.* Washington, DC: Author

Vaezghasemi, M., Eurenius, E., Ivarsson, A., Sundberg, L.R., Silfverdal, S-A., & Lindkvist, M. (2020). Socio-emotional problems among Swedish three-year-olds: An item response theory analysis of the ages and stages questionnaires: Social-emotional. *BMC Pediatrics*, 20. https://doi.org/10.1186/s12887-020-2000-y

Vitiello, V. E., & Williford, A. P. (2021). Alignment of teacher ratings and child direct assessments in preschool: A closer look at teaching strategies GOLD. *Early Childhood Research Quarterly, 56*(3), 114–123. https://doi.org/10.1016/j.ecresq.2021.03.004

Walker, C.M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, 29(4), 364-376. https://doi.org/10.1177/0734282911406666

Waterman, C., McDermott, P. A., Fantuzzo, J. W., & Gadsden, V. L. (2012). The matter of assessor variance in early childhood education—Or whose score is it anyway? *Early Childhood Research Quarterly, 27*(1), 46–54. https://doi.org/10.1016/j.ecresq.2011.06.003

Webb, M.L., Cohen, A.S., Schwanenflugel, P.J. (2008). Latent class analysis of differential item functioning on the Peabody Picture Vocabulary Test-III. *Education and Psychological Measurement*, 68(2), 335-351. https://doi.org/10.1177/0013164407308474

Widaman, K.F. & Thompson, J.S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, 8(1): 16–37. doi:10.1037/1082-989x.8.1.16.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation*, 12(3). https://doi.org/10.7275/mhqa-cd89

Zwick, R., Thayer, D. T., and Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement*, 36(1), 1-28.

Tables

Table 1. Norm sample by age/grade

Age/Grade	n	%
Birth to 1 year	5,000	15.75%
1 to 2 years	5,000	15.75%
2 to 3 years	5,000	15.75%
Preschool 3	5,000	15.75%
Prekindergarten 4	5,000	15.75%
Kindergarten	5,000	15.75%
First grade	1,250	3.94%
Second grade	375	1.18%
Third grade	126	0.4%
Total	**31,751**	**100.00%**

Table 2. Demographic characteristics

Characteristic	Category	%
Gender	Female	49.7%
	Male	50.29%
	Unknown	0.01%
Race/Ethnicity	Asian	4.16%
	African American	15.04%
	Hispanic	26.9%
	Native American and Pacific Islander	1.92%
	No Response	0.7%
	Two or More Races	5.04%
	White	46.23%
Primary Language	English	79.88%
	Other	7.46%
	Spanish	12.66%
Disability Status	Has IFSP or IEP	7.63%

Table 3. Alpha reliability coefficients by age/grade level

	Social-Emotional	Physical	Language	Cognitive	Literacy	Math
Birth to 1 year	0.90	0.89	0.89	0.91	-	-
1 to 2 years	0.91	0.88	0.93	0.93	0.88	0.85
2 to 3 years	0.94	0.92	0.95	0.95	0.92	0.90
Preschool 3	0.94	0.89	0.95	0.96	0.91	0.90
Pre-K 4	0.93	0.88	0.94	0.95	0.91	0.89
Kindergarten	0.93	0.87	0.94	0.96	0.95	0.91
1st Grade	0.91	0.89	0.97	0.97	0.95	0.91
2nd Grade	0.96	0.90	0.95	0.97	0.96	0.96
3rd Grade	0.98	0.97	0.98	0.98	0.98	0.95

Table 4. Fall item-level statistics and difficulty estimates for the social–emotional scale

						Item-Measure *r*	
Progression		**Item Difficulty**	**SE**	**Infit Mnsq**	**Outfit Mnsq**	**Point Biserial**	**Model Expected**
2a	Forms relationships with adults	-2.8	0.01	1.18	1.2	0.92	0.93
1c	Takes care of own needs appropriately	-0.85	0.01	0.94	0.93	0.95	0.95
1b	Follows limits and expectations	-0.6	0.01	0.89	0.88	0.95	0.94
1a	Manages feelings	-0.12	0.01	1.12	1.11	0.92	0.93
2b	Responds to emotional cues	0.49	0.01	0.88	0.88	0.94	0.94
2c	Interacts with peers	0.59	0.01	0.85	0.84	0.94	0.93
2d	Makes friends	0.78	0.01	0.96	1.03	0.94	0.94
3b	Solves social problems	1.21	0.01	0.96	0.95	0.95	0.95
3a	Balances needs and rights of self and others	1.29	0.01	0.86	0.83	0.95	0.94

Table 5. Fall item-level statistics and difficulty estimates for the physical scale

						Item-Measure *r*	
Progression		**Item Difficulty**	**SE**	**Infit Mnsq**	**Outfit Mnsq**	**Point Biserial**	**Model Expected**
7a	Uses fingers and hands	-1.03	0.01	0.88	0.84	0.95	0.95
4	Demonstrates traveling skills	-1.02	0.01	0.84	0.82	0.96	0.95
6	Demonstrates gross motor manipulative skills	0.38	0.01	0.84	0.81	0.95	0.94
5	Demonstrates balancing skills	0.41	0.01	0.91	0.9	0.95	0.95
7b	Uses writing and drawing tools	1.27	0.01	1.22	2.22	0.94	0.95

Table 6. Fall item-level statistics and difficulty estimates for the language scale

						Item-Measure *r*	
Progression		**Item Difficulty**	**SE**	**Infit Mnsq**	**Outfit Mnsq**	**Point Biserial**	**Model Expected**
8b	Follows directions	-2.8	0.01	1.15	1.17	0.95	0.96
9b	Speaks clearly	-0.07	0.01	0.95	0.97	0.96	0.96
8a	Comprehends language	0.01	0.01	0.9	0.9	0.96	0.95
9a	Uses an expanding expressive vocabulary	0.25	0.01	0.78	0.79	0.96	0.95
10a	Engages in conversations	0.32	0.01	0.89	0.91	0.96	0.96
9c	Uses conventional grammar	0.52	0.01	0.87	0.89	0.96	0.96
10b	Uses social rules of language	0.71	0.01	1.02	1.03	0.95	0.95
9d	Tells about another time or place	1.06	0.01	0.92	0.92	0.94	0.94

Table 7. Fall item-level statistics and difficulty estimates for the cognitive scale

Progression		Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item-Measure *r*	
						Point Biserial	Model Expected
11b	Persists	-1.6	0.01	0.95	0.96	0.95	0.95
11c	Solves problems	-0.96	0.01	0.96	0.96	0.96	0.96
11a	Attends and engages	-0.33	0.01	1.07	1.06	0.93	0.93
13	Uses classification skills	-0.06	0.01	1	1.12	0.95	0.95
14a	Thinks symbolically	-0.06	0.01	0.9	0.97	0.95	0.95
14b	Engages in sociodramatic play	0.23	0.01	1	1.04	0.95	0.95
12a	Recognizes and recalls	0.31	0.01	0.89	0.9	0.95	0.95
11d	Shows curiosity and motivation	0.55	0.01	0.89	0.91	0.95	0.95
11e	Shows flexibility and inventiveness	0.94	0.01	0.88	0.9	0.95	0.95
12b	Makes connections	0.97	0.01	0.79	0.79	0.96	0.96

Table 8. Fall item-level statistics and difficulty estimates for the literacy scale

Progression		Item Difficulty	SE	Infit Mnsq	Outfit Mnsq	Item-Measure *r*	
						Point Biserial	Model Expected
19a	Writes name	-2.88	0	1.47	1.35	0.85	0.86
15b	Notices and discriminates alliteration	-2.2	0.01	0.8	0.79	0.84	0.83
18b	Uses emergent reading skills	-2.18	0.01	0.7	0.68	0.84	0.83
16a	Identifies and names letters	-2.1	0.01	1.13	1.22	0.81	0.81
15a	Notices and discriminates rhyme	-1.68	0.01	1.02	1.05	0.85	0.86
17b	Uses print concepts	-1.19	0.01	0.7	0.77	0.83	0.82
16b	Identifies letter-sound correspondences	-1.01	0.01	0.89	1.02	0.74	0.74
17a	Uses and appreciates books and other texts	-0.32	0.01	0.86	0.95	0.9	0.9
19b	Writes to convey ideas and information	0.1	0.01	1.31	1.18	0.85	0.85
18a	Interacts during reading experiences	0.78	0.01	0.84	0.82	0.85	0.84
15c	Notices and discriminates units of sound	0.94	0.01	0.91	2.03	0.76	0.76
18c	Retells stories and recounts details	1.07	0.01	0.85	0.82	0.82	0.81
15d	Applies phonics concepts and knowledge	1.8	0.01	1.02	1.39	0.64	0.65
18e	Reads fluently	2.9	0.01	0.98	0.95	0.55	0.55
19c	Writes using conventions	2.97	0.01	1.2	1.42	0.56	0.56
18d	Uses context clues to read	2.99	0.01	1.3	4.78	0.55	0.56

Table 9. Fall item-level statistics and difficulty estimates for the mathematics scale

						Item-Measure *r*	
Progression		**Item Difficulty**	**SE**	**Infit Mnsq**	**Outfit Mnsq**	**Point Biserial**	**Model Expected**
21a	Understands spatial relationships	-4.11	0.01	1.06	1.12	0.94	0.94
20c	Connects numerals with their quantities	-0.98	0.01	0.84	0.77	0.88	0.87
20b	Quantifies	-0.92	0.01	0.69	0.71	0.92	0.91
20a	Counts	-0.87	0.01	0.78	0.79	0.94	0.93
21b	Understands shapes	-0.75	0.01	0.95	0.92	0.94	0.93
23	Demonstrates knowledge of patterns	-0.71	0.01	0.9	0.94	0.92	0.92
22a	Measures objects	-0.14	0.01	1.2	1.06	0.89	0.9
22b	Measures time and money	0.79	0.01	1.37	1.83	0.82	0.84
22c	Represents and analyzes data	1.04	0.01	1.11	1.06	0.78	0.78
20f	Applies number combinations	2.01	0.01	1.12	1.05	0.59	0.59
20d	Understands and uses place value	2.27	0.01	1.12	2.19	0.56	0.57
20e	Properties of mathematical operations	2.37	0.01	0.93	0.98	0.6	0.6

Table 10. Social–emotional scaled scores by age/grade

Age / Grade		Fall	Winter	Spring	Gain
Birth to 1 year	Mean	187.01	227.54	260.62	73.61
	SEM	19	17	17	
	SD	59.42	55.23	52.32	50.81
	25th	140	195	230	42
	50th	188.5	230	258	72
	75th	230	258	297	104
1 to 2 years	Mean	279.58	307.42	326.22	46.63
	SEM	16	15	14	
	SD	52.88	48.71	48.28	47.14
	25th	250	282	303	20
	50th	282	311	331	44
	75th	318	337	354	72
2 to 3 years	Mean	336.9	358.67	375.8	38.9
	SEM	14	13	13	
	SD	57.88	52.1	54.51	49.57
	25th	311	331	349	11
	50th	343	365	381	36
	75th	370	387	403	63
Preschool 3	Mean	385.59	417.55	440.4	54.8
	SEM	16	14	15	
	SD	63.84	59.09	61.8	51.07
	25th	360	387	409	26
	50th	392	420	442	49
	75th	420	453	473	78
Pre-K 4	Mean	425.17	463.63	491.6	66.42
	SEM	15	14	15	
	SD	55.41	52.78	58.14	49.6
	25th	398	442	463	36
	50th	431	467	499	62
	75th	457	494	522	91
Kindergarten	Mean	471.91	524.09	559.14	87.23
	SEM	17	17	17	
	SD	64.34	62.9	65.79	52.77
	25th	447	499	527	54
	50th	478	533	569	88
	75th	505	563	599	114
First Grade	Mean	553.12	608.5	640	86.88
	SEM	22	19	19	
	SD	74.96	63.28	64.29	61.59
	25th	505	587	617	46
	50th	563	617	656	81
	75th	599	644	674	128

Table 11. Physical scaled scores by age/grade

Age / Grade		Fall	Winter	Spring	Gain
Birth to 1 year	Mean	214.2	275.36	324.68	110.48
	SEM	29	26	24	
	SD	87.92	79.65	73.13	70.73
	25th	141	244	270	71
	50th	215	270	334	105
	75th	270	334	376	149
1 to 2 years	Mean	363.38	396.92	421.55	58.17
	SEM	24	23	22	
	SD	69.92	65.89	64.3	62.86
	25th	320	361	393	17
	50th	361	411	411	56
	75th	411	428	458	91
2 to 3 years	Mean	441	468.37	490.1	49.1
	SEM	21	19	20	
	SD	75.16	68.57	70.09	68.27
	25th	411	428	444	12
	50th	444	470	494	47
	75th	485	506	530	83
Preschool 3	Mean	504.22	540.96	567.01	62.79
	SEM	25	23	23	
	SD	76.26	69.32	69.67	63.29
	25th	470	506	530	25
	50th	518	543	568	59
	75th	543	581	604	89
Pre-K 4	Mean	549.21	593.38	624.23	75.02
	SEM	23	21	22	
	SD	65.37	60.18	63.83	58.32
	25th	518	568	593	41
	50th	556	604	635	74
	75th	581	625	668	105
Kindergarten	Mean	594.16	653.63	687.24	93.08
	SEM	26	23	22	
	SD	71.58	63	60.65	60.84
	25th	568	625	657	55
	50th	593	668	701	91
	75th	635	690	726	124
First Grade	Mean	687.91	723.68	746.63	58.72
	SEM	26	25	24	
	SD	77.99	74.79	72.61	48.03
	25th	668	710	734	31
	50th	701	734	757	53
	75th	726	749	771	81

Table 12. Language scaled scores by age/grade

Age / Grade		Fall	Winter	Spring	Gain
Birth to 1 year	Mean	154.82	196.47	233.94	79.13
	SEM	21	21	21	
	SD	62.95	63.49	62.25	55.28
	25th	116	157	197	44
	50th	157	197	236	80
	75th	197	236	268	111
1 to 2 years	Mean	263.38	299.68	327.33	63.95
	SEM	18	17	17	
	SD	66.55	64.21	63.84	55.23
	25th	224	268	294	31
	50th	268	302	333	62
	75th	310	341	364	93
2 to 3 years	Mean	349.2	377.32	400.73	51.53
	SEM	17	16	17	
	SD	76.96	73.31	74.08	58.25
	25th	318	341	364	21
	50th	356	380	410	47
	75th	395	424	445	78
Preschool 3	Mean	409.23	444.55	472.36	63.13
	SEM	19	19	19	
	SD	86.21	83.19	85.2	59.96
	25th	372	410	431	28
	50th	417	452	481	58
	75th	459	491	515	90
Pre-K 4	Mean	456.98	500.6	535.8	78.82
	SEM	18	18	20	
	SD	74.36	75.05	80.52	58.5
	25th	424	466	497	42
	50th	466	506	542	76
	75th	497	542	585	112
Kindergarten	Mean	511.32	571.54	610.17	98.86
	SEM	22	22	22	
	SD	90.09	87.85	88.1	62.06
	25th	481	542	576	60
	50th	524	585	627	96
	75th	560	619	659	134
First Grade	Mean	590.21	663.22	699.15	108.95
	SEM	17	15	14	
	SD	100.1	85.47	79.11	75.64
	25th	533	644	679	55
	50th	602	679	718	98.5
	75th	659	707	746	155

Table 13. Cognitive scaled scores by age/grade

Age / Grade		Fall	Winter	Spring	Gain
Birth to 1 year	Mean	150.44	192.45	226.81	76.37
	SEM	19	17	16	
	SD	63.22	56.98	54.39	52.09
	25th	116	159	196	43
	50th	159	196	224	74
	75th	196	224	262	108
1 to 2 years	Mean	251.96	283.94	307.16	55.2
	SEM	15	14	14	
	SD	56.32	53.48	53.79	49.02
	25th	215	255	276	28
	50th	255	291	312	52
	75th	291	319	338	81
2 to 3 years	Mean	323.58	350.02	371.37	47.79
	SEM	14	13	14	
	SD	62.98	59.17	61.23	54.48
	25th	298	319	338	18
	50th	326	357	375	45
	75th	363	388	401	75
Preschool 3	Mean	385.29	419.83	444.9	59.61
	SEM	14	13	13	
	SD	67.97	65.15	67.06	54.48
	25th	357	388	408	27
	50th	388	422	448	55
	75th	429	459	481	85
Pre-K 4	Mean	429.08	471.85	503.01	73.94
	SEM	13	13	14	
	SD	59.96	58.45	63.44	51.24
	25th	401	448	470	41
	50th	442	481	509	71
	75th	465	509	545	103
Kindergarten	Mean	482.18	543.65	582.45	100.27
	SEM	15	14	14	
	SD	74.46	68.93	70.82	59.47
	25th	454	514	545	63
	50th	492	557	597	101
	75th	521	588	630	129
First Grade	Mean	572.27	640.79	670.68	98.41
	SEM	15	11	11	
	SD	88.07	66.24	62.67	72.71
	25th	521	630	651	53
	50th	597	650	683	85.5
	75th	634	675	709	136

Table 14. Literacy scaled scores by age/grade

Age / Grade		Fall	Winter	Spring	Gain
Birth to 1 year	Mean	115.68	179.69	237.21	121.53
	SEM	73	76	72	
	SD	102.95	107.85	102.02	102.16
	25th	0	110	207	47
	50th	110	207	271	110
	75th	207	271	305	207
1 to 2 years	Mean	285.66	324.72	350.44	64.78
	SEM	25	22	21	
	SD	89.85	77.9	75.19	78.83
	25th	246	290	318	23
	50th	305	341	360	55
	75th	341	369	392	95.5
2 to 3 years	Mean	373.07	397.19	416.48	43.41
	SEM	22	20	19	
	SD	74.63	67.38	62.96	59.99
	25th	351	375	392	15
	50th	385	404	425	37
	75th	415	439	450	65
Preschool 3	Mean	433.83	463.16	480.54	46.7
	SEM	20	16	16	
	SD	66.88	54.98	51.85	49.77
	25th	410	445.25	460	22
	50th	446	470	487	40
	75th	472	493	508	63
Pre-K 4	Mean	473.06	504.93	524.49	51.43
	SEM	16	13	14	
	SD	51.85	44.49	46.47	40.6
	25th	457	489	506	31
	50th	482	509	529	48
	75th	502	531	550	67
Kindergarten	Mean	523.46	566.6	596.6	73.13
	SEM	12	13	13	
	SD	53.95	56.08	58.89	39.42
	25th	508	545	568	47
	50th	529	571	604	72
	75th	548	598	632	96
First Grade	Mean	600.86	650.28	679.05	78.19
	SEM	13	12	13	
	SD	57.63	54.73	56.63	48.15
	25th	564	627	657	46
	50th	609	657	689	74
	75th	637	679	714	107

Table 15. Mathematics scaled scores by age/grade

Age / Grade		Fall	Winter	Spring	Gain
Birth to 1 year	Mean	17.48	43.41	79.18	61.7
	SEM	30	44	53	
	SD	42.74	61.82	74.96	67.66
	25th	0	0	0	0
	50th	0	0	56	56
	75th	0	93	136	117
1 to 2 years	Mean	124.05	166.01	198.01	73.97
	SEM	30	29	27	
	SD	77.53	74.75	70.87	67.68
	25th	56	117	167	31
	50th	136	181	206	70
	75th	181	218	239	113
2 to 3 years	Mean	225.12	252.96	276.12	51
	SEM	23	22	21	
	SD	72.24	68.22	66.94	56.77
	25th	194	229	249	20
	50th	239	261	286	47
	75th	273	292	315	76
Preschool 3	Mean	296.98	329.97	351.06	54.09
	SEM	22	19	19	
	SD	68.77	61.31	60.05	48.86
	25th	273	304	325	27
	50th	304	335	354	49
	75th	335	367	384	74
Pre-K 4	Mean	341.81	378.8	403.41	61.6
	SEM	19	17	18	
	SD	56.91	52.53	54.1	43.22
	25th	320	359	380	37
	50th	350	384	408	57
	75th	376	408	437	82
Kindergarten	Mean	395.61	447.16	483.08	87.47
	SEM	19	18	19	
	SD	64.16	59.41	63.36	48.94
	25th	371	425	454	57
	50th	404	458	490	86
	75th	429	483	519	114
First Grade	Mean	473.8	545.4	585.78	111.98
	SEM	22	19	18	
	SD	73.5	61.94	59.83	64.42
	25th	429	519	566	74
	50th	483	552	599	105
	75th	519	580	622	148

Table 16. Widely held expectations for social–emotional by age/grade

Age / Grade	Level	Fall	Winter	Spring
Birth to 1 year	Below Expectations	5.1	1.1	0.5
	Meets Expectations	86	74.1	50
	Exceeds Expectations	8.9	24.8	49.4
1 to 2 years	Below Expectations	15.9	7	3.9
	Meets Expectations	79.8	80.7	71.4
	Exceeds Expectations	4.3	12.4	24.7
2 to 3 years	Below Expectations	26.7	15.6	9.8
	Meets Expectations	66.1	71	65.8
	Exceeds Expectations	7.1	13.4	24.4
Preschool 3	Below Expectations	34.2	16.1	9.5
	Meets Expectations	59.1	67.4	60.4
	Exceeds Expectations	6.7	16.5	30.1
Pre-K 4	Below Expectations	55.4	24.9	13.7
	Meets Expectations	42.9	67.7	62.9
	Exceeds Expectations	1.7	7.3	23.4
Kindergarten	Below Expectations	60.6	20.4	11.8
	Meets Expectations	37.8	75.5	69.6
	Exceeds Expectations	1.6	4.1	18.6
First Grade	Below Expectations	66.6	30.9	13.5
	Meets Expectations	31	63.4	71.8
	Exceeds Expectations	2.4	5.8	14.7

Table 17. Widely held expectations for physical by age/grade

Age / Grade	Level	Fall	Winter	Spring
Birth to 1 year	Below Expectations	13.1	2.9	0.8
	Meets Expectations	79.3	71.6	48.4
	Exceeds Expectations	7.6	25.5	50.8
1 to 2 years	Below Expectations	18.6	8.5	4.9
	Meets Expectations	70.4	67.1	55.3
	Exceeds Expectations	11	24.5	39.8
2 to 3 years	Below Expectations	20.8	11.6	7.1
	Meets Expectations	69.3	70.5	63.4
	Exceeds Expectations	9.9	17.9	29.5
Preschool 3	Below Expectations	26.4	12.6	6.7
	Meets Expectations	67.3	71	63.8
	Exceeds Expectations	6.3	16.4	29.5
Pre-K 4	Below Expectations	41.5	16	8.4
	Meets Expectations	57.5	79.1	75.5
	Exceeds Expectations	1	4.9	16
Kindergarten	Below Expectations	60.4	19.1	10.4
	Meets Expectations	37.6	76.2	73.1
	Exceeds Expectations	2	4.7	16.5
First Grade	Below Expectations	50.1	17.6	8.9
	Meets Expectations	46.6	74.2	73.5
	Exceeds Expectations	3.4	8.2	17.6

Table 18. Widely held expectations for language by age/grade

Age / Grade	Level	Fall	Winter	Spring
Birth to 1 year	Below Expectations	12	4.1	1.8
	Meets Expectations	85.4	83.5	67
	Exceeds Expectations	2.7	12.4	31.2
1 to 2 years	Below Expectations	32.6	15.2	8.9
	Meets Expectations	65.1	76.4	71.7
	Exceeds Expectations	2.4	8.4	19.4
2 to 3 years	Below Expectations	34.9	21.7	14.3
	Meets Expectations	58.9	66.2	63.9
	Exceeds Expectations	6.2	12.1	21.7
Preschool 3	Below Expectations	40.6	24	15.8
	Meets Expectations	55.1	65.3	63.5
	Exceeds Expectations	4.2	10.7	20.7
Pre-K 4	Below Expectations	50.4	25.1	15.8
	Meets Expectations	48.5	70.5	68.8
	Exceeds Expectations	1.1	4.4	15.4
Kindergarten	Below Expectations	68.9	28.3	17.2
	Meets Expectations	28.8	66	61.8
	Exceeds Expectations	2.2	5.6	21
First Grade	Below Expectations	69.8	31.3	14.9
	Meets Expectations	27.4	63.7	73.5
	Exceeds Expectations	2.9	5	11.6

Table 19. Widely held expectations for cognitive by age/grade

Age / Grade	Level	Fall	Winter	Spring
Birth to 1 year	Below Expectations	8.3	1.8	0.5
	Meets Expectations	83.5	75	51.2
	Exceeds Expectations	8.1	23.3	48.3
1 to 2 years	Below Expectations	8.4	2.8	1.6
	Meets Expectations	85.7	78.8	65
	Exceeds Expectations	6	18.4	33.4
2 to 3 years	Below Expectations	24.8	13.8	8.5
	Meets Expectations	68.8	73	66.9
	Exceeds Expectations	6.4	13.3	24.6
Preschool 3	Below Expectations	39	20.1	12.2
	Meets Expectations	56.3	66.8	63.8
	Exceeds Expectations	4.6	13.1	24
Pre-K 4	Below Expectations	54.7	24.4	13.4
	Meets Expectations	44.7	72.7	74.5
	Exceeds Expectations	0.6	3	12.1
Kindergarten	Below Expectations	68.4	25.8	14.1
	Meets Expectations	30.3	72.3	76.8
	Exceeds Expectations	1.3	1.9	9.1
First Grade	Below Expectations	78	34.9	16.8
	Meets Expectations	20.6	61.6	71
	Exceeds Expectations	1.4	3.5	12.2

Table 20. Widely held expectations for literacy by age/grade

Age / Grade	Level	Fall	Winter	Spring
Birth to 1 year	Below Expectations	0	0	0
	Meets Expectations	84.5	63.5	39.6
	Exceeds Expectations	15.5	36.5	60.4
1 to 2 years	Below Expectations	10	4.1	2.6
	Meets Expectations	82.6	77.5	62.5
	Exceeds Expectations	7.5	18.5	34.9
2 to 3 years	Below Expectations	33.8	19.9	12.4
	Meets Expectations	60.6	68.7	66.4
	Exceeds Expectations	5.6	11.4	21.3
Preschool 3	Below Expectations	48.5	25	15
	Meets Expectations	47	62.6	61.9
	Exceeds Expectations	4.5	12.4	23.1
Pre-K 4	Below Expectations	54.2	21.9	11.4
	Meets Expectations	45.1	74.2	74.5
	Exceeds Expectations	0.7	3.8	14
Kindergarten	Below Expectations	53.3	17.1	10
	Meets Expectations	45.5	78.4	69.1
	Exceeds Expectations	1.2	4.5	20.9
First Grade	Below Expectations	79.3	38.2	19.3
	Meets Expectations	19.7	58.2	68.5
	Exceeds Expectations	1	3.6	12.2

Table 21. Widely held expectations for mathematics by age/grade

Age / Grade	Level	Fall	Winter	Spring
Birth to 1 year	Below Expectations	0	0	0
	Meets Expectations	95.1	83.7	63
	Exceeds Expectations	4.9	16.3	37
1 to 2 years	Below Expectations	27.6	12.6	6.4
	Meets Expectations	68.6	75.4	68.6
	Exceeds Expectations	3.8	12	25
2 to 3 years	Below Expectations	33.5	18.9	12.2
	Meets Expectations	62.1	70.9	67.9
	Exceeds Expectations	4.4	10.1	19.9
Preschool 3	Below Expectations	40.3	18.9	11.1
	Meets Expectations	52.8	64.2	59.1
	Exceeds Expectations	6.9	17	29.8
Pre-K 4	Below Expectations	68.5	34.6	18.3
	Meets Expectations	30.7	61.1	66.8
	Exceeds Expectations	0.9	4.3	14.9
Kindergarten	Below Expectations	71.6	26.7	13.6
	Meets Expectations	26.9	70	67.6
	Exceeds Expectations	1.5	3.2	18.8
First Grade	Below Expectations	79.9	29.9	11
	Meets Expectations	18.7	65.3	80.2
	Exceeds Expectations	1.4	4.8	8.8

Table 22. Alpha Reliability coefficients by age/grade level and gender

Age / Grade	Gender	Social-	Physical		Cognitive	Literacy	Math
Birth to 1 year	Female	0.90	0.88	0.89	0.90	-	-
	Male	0.90	0.91	0.89	0.91	-	-
1 to 2 years	Female	0.92	0.89	0.93	0.93	0.87	0.84
	Male	0.90	0.87	0.93	0.93	0.89	0.86
2 to 3 years	Female	0.93	0.92	0.95	0.95	0.92	0.90
	Male	0.94	0.92	0.95	0.95	0.92	0.89
Preschool 3	Female	0.93	0.90	0.94	0.95	0.91	0.90
	Male	0.94	0.89	0.95	0.96	0.91	0.89
Pre-K 4	Female	0.93	0.88	0.93	0.94	0.90	0.89
	Male	0.94	0.89	0.94	0.95	0.91	0.90
Kindergarten	Female	0.92	0.86	0.94	0.96	0.95	0.90
	Male	0.94	0.87	0.95	0.96	0.95	0.91

Table 23. Alpha Reliability coefficients by age/grade level and ethnicity

Age / Grade	Ethnicity	Social-Emotional	Physical	Language	Cognitive	Literacy	Math
Birth to 1 year	Black NH	0.89	0.85	0.90	0.92	-	-
	Hispanic	0.88	0.88	0.86	0.88	-	-
	White NH	0.90	0.91	0.89	0.91	-	-
1 to 2 years	Black NH	0.91	0.88	0.92	0.93	0.91	0.87
	Hispanic	0.92	0.88	0.92	0.92	0.84	0.82
	White NH	0.91	0.88	0.93	0.93	0.87	0.85
2 to 3 years	Black NH	0.94	0.93	0.95	0.96	0.92	0.90
	Hispanic	0.93	0.92	0.95	0.95	0.91	0.88
	White NH	0.94	0.91	0.95	0.95	0.93	0.91
Preschool 3	Black NH	0.94	0.91	0.95	0.96	0.92	0.90
	Hispanic	0.94	0.89	0.95	0.96	0.90	0.89
	White NH	0.93	0.88	0.95	0.96	0.91	0.90
Pre-K 4	Black NH	0.94	0.91	0.93	0.95	0.90	0.90
	Hispanic	0.94	0.89	0.93	0.95	0.90	0.89
	White NH	0.93	0.87	0.94	0.94	0.91	0.89
Kindergarten	Black NH	0.94	0.85	0.92	0.95	0.95	0.90
	Hispanic	0.93	0.85	0.95	0.96	0.95	0.90
	White NH	0.93	0.87	0.95	0.96	0.95	0.91

Table 24. Alpha Reliability coefficients by age/grade level and primary language

Age / Grade	Language	Social-Emotional	Physical	Language	Cognitive	Literacy	Math
Birth to 1 year	English	0.90	0.90	0.90	0.91	-	-
	Spanish	0.88	0.87	0.86	0.87	-	-
1 to 2 years	English	0.91	0.88	0.93	0.93	0.88	0.85
	Spanish	0.91	0.88	0.93	0.93	0.85	0.83
2 to 3 years	English	0.94	0.92	0.95	0.95	0.92	0.90
	Spanish	0.93	0.92	0.95	0.95	0.91	0.88
Preschool 3	English	0.94	0.89	0.95	0.96	0.91	0.90
	Spanish	0.93	0.89	0.95	0.95	0.90	0.89
Pre-K 4	English	0.93	0.88	0.94	0.95	0.91	0.89
	Spanish	0.94	0.89	0.94	0.96	0.90	0.89
Kindergarten	English	0.93	0.86	0.94	0.96	0.95	0.91
	Spanish	0.93	0.88	0.95	0.96	0.95	0.90

Table 25. Alpha Reliability coefficients by age/grade level and disability status

Age / Grade	IEP/IFSP	Emotional	Physical	Language	Cognitive	Literacy	Math
1 to 2 years	No	0.91	0.89	0.93	0.93	0.88	0.85
	Yes	0.92	0.87	0.93	0.93	0.92	0.88
2 to 3 years	No	0.94	0.92	0.95	0.95	0.92	0.90
	Yes	0.92	0.91	0.94	0.94	0.88	0.85
Preschool 3	No	0.93	0.89	0.94	0.95	0.91	0.89
	Yes	0.94	0.89	0.96	0.96	0.91	0.91
Pre-K 4	No	0.93	0.88	0.93	0.94	0.90	0.88
	Yes	0.94	0.88	0.95	0.96	0.91	0.91
Kindergarten	No	0.93	0.87	0.94	0.95	0.95	0.90
	Yes	0.96	0.86	0.96	0.97	0.95	0.94

Birth to 1 year comparison was omitted due to insufficient sample size

Table 26. Alpha Reliability coefficients by age/grade level and free-reduced lunch status

Age / Grade	FRL	Social-Emotional	Physical	Language	Cognitive	Literacy	Math
Preschool 3	No	0.94	0.89	0.95	0.95	0.93	0.92
	Yes	0.93	0.89	0.93	0.95	0.93	0.93
Pre-K 4	No	0.93	0.88	0.93	0.94	0.94	0.89
	Yes	0.94	0.88	0.93	0.95	0.90	0.92
Kindergarten	No	0.94	0.87	0.94	0.96	0.92	0.89
	Yes	0.93	0.86	0.93	0.95	0.98	0.88

Table 27. Differential item functioning results by gender for the social–emotional domain

Progression		Female Rasch Item Difficulty	Male Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
1a	-0.12	Average	-0.12	Average	0.00	Negligible	0.02	0.00		1.09	
1b	-0.63	Easy	-0.56	Easy	-0.07	Negligible	0.02	-4.44	***	22.90	***
1c	-0.85	Easy	-0.85	Easy	0.00	Negligible	0.01	0.00		1.54	
2a	-2.75	Easy	-2.85	Easy	0.11	Negligible	0.02	6.68	***	33.36	***
2b	0.49	Average	0.49	Average	0.00	Negligible	0.02	0.00		5.25	*
2c	0.59	Difficult	0.59	Difficult	0.00	Negligible	0.01	0.00		2.13	
2d	0.81	Difficult	0.76	Difficult	0.05	Negligible	0.01	3.32	***	11.82	***
3a	1.29	Difficult	1.29	Difficult	0.00	Negligible	0.01	0.00		2.71	
3b	1.21	Difficult	1.21	Difficult	0.00	Negligible	0.02	0.00		0.11	

Note. ***** = p<.001, ** = p<.01, * = p<.05.**

Table 28. Differential item functioning results by gender for the physical domain

Progression		Female Rasch Item Difficulty	Male Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
4	-0.97	Easy	-1.08	Easy	0.10	Negligible	0.02	5.99	***	40.62	***
5	0.44	Average	0.38	Average	0.06	Negligible	0.02	3.65	***	15.67	***
6	0.51	Difficult	0.26	Average	0.25	Negligible	0.02	14.32	***	100.00	***
7a	-1.11	Easy	-0.96	Easy	-0.15	Negligible	0.02	-8.55	***	82.44	***
7b	1.15	Difficult	1.38	Difficult	-0.23	Negligible	0.02	-14.60	***	100.00	***

Note. ***** = *p*<.001, ** = *p*<.01, * = *p*<.05.**

Table 29. Differential item functioning results by gender for the language domain

Progression		Female Rasch Item Difficulty	Male Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
8a	0.04	Average	-0.01	Average	0.05	Negligible	0.02	2.43	*	7.56	**
8b	-2.80	Easy	-2.78	Easy	-0.02	Negligible	0.02	-1.16		4.27	*
9a	0.25	Average	0.25	Average	0.00	Negligible	0.02	0.00		1.08	
9b	-0.10	Average	-0.04	Average	-0.05	Negligible	0.02	-2.98	**	11.07	***
9c	0.52	Difficult	0.52	Difficult	0.00	Negligible	0.02	0.00		1.66	
9d	1.09	Difficult	1.04	Difficult	0.05	Negligible	0.02	2.89	**	8.50	**
10a	0.32	Average	0.32	Average	0.00	Negligible	0.02	0.00		3.50	
10b	0.71	Difficult	0.71	Difficult	0.00	Negligible	0.02	0.00		0.08	

Note. ***** = *p*<.001, ** = *p*<.01, * = *p*<.05.**

Table 30. Differential item functioning results by gender for the cognitive domain

Progression		Female Rasch Item Difficulty	Male Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint se	Rasch-Welch t	p	Mantel-Haenszel x^2	p
11a	-0.39	Average	-0.26	Average	-0.13	Negligible	0.02	-6.95	***	40.69	***
11b	-1.63	Easy	-1.57	Easy	-0.06	Negligible	0.02	-3.06	**	10.19	**
11c	-0.96	Easy	-0.96	Easy	0.00	Negligible	0.02	0.00		0.30	
11d	0.59	Difficult	0.52	Difficult	0.07	Negligible	0.02	3.63	***	14.55	***
11e	0.96	Difficult	0.94	Difficult	0.02	Negligible	0.02	1.32	*	5.74	*
12a	0.31	Average	0.31	Average	0.00	Negligible	0.02	0.00		2.97	
12b	0.97	Difficult	0.97	Difficult	0.00	Negligible	0.02	0.00		3.05	
13	0.00	Average	-0.11	Average	0.11	Negligible	0.02	5.90	***	34.11	***
14a	-0.06	Average	-0.06	Average	0.00	Negligible	0.02	0.00		2.01	
14b	0.18	Average	0.28	Average	-0.10	Negligible	0.02	-5.77	***	39.41	***

NOTE. ***** = *P*<.001, ** = *P*<.01, * = *P*<.05.**

Table 31. Differential item functioning results by gender for the literacy domain

Progression		Female Rasch Item Difficulty	Male Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
15a	-1.71	Easy	-1.66	Easy	-0.05	Negligible	0.01	-3.77	***	13.73	***
15b	-2.20	Easy	-2.20	Easy	0.00	Negligible	0.01	0.00		5.22	*
15c	0.94	Difficult	0.94	Difficult	0.00	Negligible	0.01	0.00		3.99	*
15d	1.82	Difficult	1.80	Difficult	0.02	Negligible	0.02	1.22		3.84	
16a	-2.04	Easy	-2.16	Easy	0.12	Negligible	0.01	9.74	***	76.60	***
16b	-0.97	Easy	-1.03	Easy	0.06	Negligible	0.01	4.09	***	17.82	***
17a	-0.32	Average	-0.32	Average	0.00	Negligible	0.02	0.00		1.18	
17b	-1.19	Easy	-1.19	Easy	0.00	Negligible	0.01	0.00		0.06	
18a	0.78	Difficult	0.78	Difficult	0.00	Negligible	0.01	0.00		0.53	
18b	-2.18	Easy	-2.18	Easy	0.00	Negligible	0.01	0.00		10.01	**
18c	1.07	Difficult	1.07	Difficult	0.00	Negligible	0.01	0.00		5.40	*
18d	3.04	Difficult	2.95	Difficult	0.09	Negligible	0.03	3.50	***	6.58	*
18e	2.93	Difficult	2.88	Difficult	0.05	Negligible	0.03	2.04	*	3.31	
19a	-2.88	Easy	-2.88	Easy	0.00	Negligible	0.01	0.00		2.85	
19b	0.08	Average	0.13	Average	-0.05	Negligible	0.01	-4.39	***	13.68	***
19c	2.97	Difficult	2.97	Difficult	0.00	Negligible	0.03	0.00		0.02	

Note. ****** = p<.001, ** = p<.01, * = p<.05.***

Table 32. Differential item functioning results by gender for the mathematics domain

Progression		Female Rasch Item Difficulty	Male Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
20a	-0.87	Easy	-0.87	Easy	0.00	Negligible	0.01	0.00		3.44	
20b	-0.92	Easy	-0.92	Easy	0.00	Negligible	0.01	0.00		3.05	
20c	-0.94	Easy	-1.01	Easy	0.07	Negligible	0.01	5.47	***	37.62	***
20d	2.32	Difficult	2.23	Difficult	0.09	Negligible	0.03	3.21	**	6.79	**
20e	2.41	Difficult	2.34	Difficult	0.07	Negligible	0.03	2.67	**	6.31	*
20f	2.06	Difficult	1.96	Difficult	0.09	Negligible	0.02	3.75	***	9.69	**
21a	-4.15	Easy	-4.08	Easy	-0.07	Negligible	0.01	-4.48	***	17.50	***
21b	-0.75	Easy	-0.75	Easy	0.00	Negligible	0.01	0.00		0.23	
22a	-0.14	Average	-0.14	Average	0.00	Negligible	0.01	0.00		1.82	
22b	0.79	Difficult	0.79	Difficult	0.00	Negligible	0.02	0.00		1.74	
22c	1.04	Difficult	1.04	Difficult	0.00	Negligible	0.02	0.00		0.04	
23	-0.71	Easy	-0.71	Easy	0.00	Negligible	0.01	0.00		1.56	

Note. ****** = p<.001, ** = p<.01, * = p<.05.***

Table 33. Differential item functioning results by primary language for the social–emotional domain

Progression		English Rasch Item Difficulty	Spanish Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
1a	-0.12	Average	-0.14	Average	0.02	Negligible	0.02	0.96		0.50	
1b	-0.60	Easy	-0.60	Easy	0.00	Negligible	0.02	0.00		0.54	
1c	-0.85	Easy	-0.90	Easy	0.05	Negligible	0.02	2.19	*	11.49	***
2a	-2.83	Easy	-2.67	Easy	-0.15	Negligible	0.02	-6.91	***	49.23	***
2b	0.49	Average	0.51	Difficult	-0.02	Negligible	0.02	-0.96		0.66	
2c	0.59	Difficult	0.72	Difficult	-0.13	Negligible	0.02	-6.04	***	50.63	***
2d	0.78	Difficult	0.87	Difficult	-0.08	Negligible	0.02	-3.88	***	21.52	***
3a	1.31	Difficult	1.19	Difficult	0.13	Negligible	0.02	6.03	***	47.33	***
3b	1.24	Difficult	1.06	Difficult	0.18	Negligible	0.02	8.21	***	74.64	***

Note. ***** = p<.001, ** = p<.01, * = p<.05.**

Table 34. Differential item functioning results by primary language for the physical domain

Progression		English Rasch Item Difficulty	Spanish Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel χ^2	*p*
4	-1.02	Easy	-0.98	Easy	-0.04	Negligible	0.03	-1.61		3.32	
5	0.43	Average	0.26	Average	0.18	Negligible	0.02	7.15	***	61.19	***
6	0.38	Average	0.44	Average	-0.05	Negligible	0.03	-2.07	*	5.96	*
7a	-1.03	Easy	-1.08	Easy	0.04	Negligible	0.03	1.60		5.27	*
7b	1.27	Difficult	1.36	Difficult	-0.10	Negligible	0.02	-4.10	***	23.60	***

Note. ***** = *p*<.001, ** = *p*<.01, * = *p*<.05.**

Table 35. Differential item functioning results by primary language for the language domain

Progression		English Rasch Item Difficulty	Spanish Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel χ^2	*p*
8a	0.01	Average	0.04	Average	-0.02	Negligible	0.03	-0.81		0.81	
8b	-2.77	Easy	-3.00	Easy	0.23	Negligible	0.03	8.28	***	54.90	***
9a	0.25	Average	0.29	Average	-0.04	Negligible	0.03	-1.34		0.42	
9b	-0.09	Average	0.11	Average	-0.20	Negligible	0.03	-7.11	***	59.30	***
9c	0.52	Difficult	0.68	Difficult	-0.17	Negligible	0.03	-6.35	***	64.51	***
9d	1.06	Difficult	1.02	Difficult	0.04	Negligible	0.03	1.67		5.71	*
10a	0.32	Average	0.35	Average	-0.02	Negligible	0.03	-0.95		0.12	
10b	0.73	Difficult	0.55	Difficult	0.18	Negligible	0.03	6.73	***	41.92	***

Note. ***** = *p*<.001, ** = *p*<.01, * = *p*<.05.**

Table 36. Differential item functioning results by primary language for the cognitive domain

Progression		English Rasch Item Difficulty	Spanish Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint se	Rasch-Welch t	*p*	Mantel-Haenszel x^2	*p*
11a	-0.33	Average	-0.27	Average	-0.06	Negligible	0.03	-2.11	*	8.77	**
11b	-1.58	Easy	-1.75	Easy	0.18	Negligible	0.03	6.63	***	48.33	***
11c	-0.92	Easy	-1.19	Easy	0.27	Negligible	0.03	10.54	***	100.00	***
11d	0.55	Difficult	0.65	Difficult	-0.10	Negligible	0.03	-3.74	***	17.38	***
11e	0.94	Difficult	0.99	Difficult	-0.05	Negligible	0.02	-1.96		4.72	*
12a	0.31	Average	0.37	Average	-0.06	Negligible	0.03	-2.37	*	7.85	**
12b	0.97	Difficult	1.08	Difficult	-0.11	Negligible	0.03	-4.36	***	28.07	***
13	-0.06	Average	-0.02	Average	-0.04	Negligible	0.03	-1.44		2.91	
14a	-0.06	Average	-0.10	Average	0.05	Negligible	0.03	1.74		5.71	*
14b	0.23	Average	0.28	Average	-0.05	Negligible	0.03	-1.80		3.83	

NOTE. ***** = *P*<.001, ** = *P*<.01, * = *P*<.05.**

Table 37. Differential item functioning results by primary language for the literacy domain

Progression		English Rasch Item Difficulty	Spanish Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
15a	-1.68	Easy	-1.55	Easy	-0.14	Negligible	0.02	-5.71	***	36.02	***
15b	-2.20	Easy	-2.15	Easy	-0.05	Negligible	0.02	-2.04	*	3.13	
15c	0.94	Difficult	0.94	Difficult	0.00	Negligible	0.02	0.00		0.03	
15d	1.80	Difficult	1.71	Difficult	0.09	Negligible	0.03	2.51	*	5.06	*
16a	-2.10	Easy	-1.97	Easy	-0.13	Negligible	0.02	-6.21	***	41.19	***
16b	-1.01	Easy	-0.96	Easy	-0.05	Negligible	0.02	-1.98	*	4.61	*
17a	-0.32	Average	-0.22	Average	-0.10	Negligible	0.03	-3.92	***	16.59	***
17b	-1.19	Easy	-1.19	Easy	0.00	Negligible	0.02	0.00		2.05	
18a	0.78	Difficult	0.90	Difficult	-0.12	Negligible	0.02	-5.20	***	24.70	***
18b	-2.18	Easy	-2.18	Easy	0.00	Negligible	0.02	0.00		0.09	
18c	1.07	Difficult	1.11	Difficult	-0.05	Negligible	0.02	-1.98	*	5.00	*
18d	2.99	Difficult	2.83	Difficult	0.16	Negligible	0.05	3.17	**	1.06	
18e	2.90	Difficult	2.74	Difficult	0.16	Negligible	0.05	3.13	**	9.82	**
19a	-2.88	Easy	-2.97	Easy	0.09	Negligible	0.02	5.92	***	22.62	***
19b	0.12	Average	-0.06	Average	0.18	Negligible	0.02	10.14	***	75.36	***
19c	2.99	Difficult	2.76	Difficult	0.22	Negligible	0.05	4.56	***	10.28	**

Note. ***** = p<.001, ** = p<.01, * = p<.05.**

Table 38. Differential item functioning results by primary language for the mathematics domain

Progression		English Rasch Item Difficulty	Spanish Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
20a	-0.87	Easy	-0.87	Easy	0.00	Negligible	0.02	0.00		0.64	
20b	-0.92	Easy	-0.98	Easy	0.06	Negligible	0.02	2.87	**	30.01	***
20c	-0.98	Easy	-0.98	Easy	0.00	Negligible	0.02	0.00		0.36	
20d	2.27	Difficult	2.20	Difficult	0.07	Negligible	0.05	1.33		0.12	
20e	2.37	Difficult	2.24	Difficult	0.13	Negligible	0.05	2.83	**	7.96	**
20f	2.01	Difficult	1.92	Difficult	0.09	Negligible	0.04	1.92		0.58	
21a	-4.11	Easy	-4.04	Easy	-0.07	Negligible	0.02	-3.29	***	17.97	***
21b	-0.75	Easy	-0.67	Easy	-0.08	Negligible	0.02	-3.86	***	22.14	***
22a	-0.14	Average	-0.25	Average	0.11	Negligible	0.02	5.09	***	34.37	***
22b	0.79	Difficult	0.89	Difficult	-0.10	Negligible	0.02	-4.10	***	29.65	***
22c	1.04	Difficult	1.01	Difficult	0.03	Negligible	0.03	1.11		0.43	
23	-0.71	Easy	-0.71	Easy	0.00	Negligible	0.02	0.00		1.57	

Note. ***** = p<.001, ** = p<.01, * = p<.05.**

Table 39. Differential item functioning results by race/ethnicity for the social–emotional domain

Progression		White Rasch Item Difficulty	Non-white Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
1a	-0.12	Average	-0.10	Average	-0.03	Negligible	0.02	-1.60		9.25	**
1b	-0.63	Easy	-0.56	Easy	-0.07	Negligible	0.02	-4.77	***	24.52	***
1c	-0.88	Easy	-0.81	Easy	-0.08	Negligible	0.01	-5.10	***	25.96	***
2a	-2.90	Easy	-2.70	Easy	-0.21	Negligible	0.02	-12.70	***	100.00	***
2b	0.49	Average	0.49	Average	0.00	Negligible	0.02	0.00		0.02	
2c	0.59	Difficult	0.59	Difficult	0.00	Negligible	0.01	0.00		8.31	**
2d	0.82	Difficult	0.74	Difficult	0.08	Negligible	0.01	5.40	***	32.29	***
3a	1.36	Difficult	1.22	Difficult	0.14	Negligible	0.01	9.73	***	100.00	***
3b	1.26	Difficult	1.16	Difficult	0.10	Negligible	0.02	6.32	***	41.39	***

Note. ***** = p<.001, ** = p<.01, * = p<.05.**

Table 40. Differential item functioning results by race/ethnicity for the physical domain

Progression		White Rasch Item Difficulty	Non-white Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
4	-1.02	Easy	-1.02	Easy	0.00	Negligible	0.02	0.00		1.59	
5	0.41	Average	0.41	Average	0.00	Negligible	0.02	0.00		5.84	*
6	0.38	Average	0.38	Average	0.00	Negligible	0.02	0.00		5.98	*
7a	-1.03	Easy	-1.03	Easy	0.00	Negligible	0.02	0.00		0.48	
7b	1.24	Difficult	1.30	Difficult	-0.06	Negligible	0.02	-3.44	***	18.62	***

Note. ***** = *p*<.001, ** = *p*<.01, * = *p*<.05.**

Table 41. Differential item functioning results by race/ethnicity for the language domain

Progression		White Rasch Item Difficulty	Non-white Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
8a	-0.02	Average	0.05	Average	-0.07	Negligible	0.02	-3.59	***	14.85	***
8b	-2.77	Easy	-2.84	Easy	0.07	Negligible	0.02	3.71	**	8.37	**
9a	0.21	Average	0.32	Average	-0.11	Negligible	0.02	-5.87	***	33.92	***
9b	-0.11	Average	-0.02	Average	-0.09	Negligible	0.02	-4.68	***	25.58	***
9c	0.52	Difficult	0.55	Difficult	-0.03	Negligible	0.02	-1.70	*	5.06	*
9d	1.09	Difficult	1.03	Difficult	0.06	Negligible	0.02	3.51	***	13.43	***
10a	0.32	Average	0.32	Average	0.00	Negligible	0.02	0.00		0.45	
10b	0.77	Difficult	0.63	Difficult	0.14	Negligible	0.02	8.08	***	58.85	***

Note. ***** = *p*<.001, ** = *p*<.01, * = *p*<.05.**

Table 42. Differential item functioning results by race/ethnicity for the cognitive domain

Progression		White Rasch Item Difficulty	Non-white Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
11a	-0.37	Average	-0.27	Average	-0.10	Negligible	0.02	-5.17	***	33.33	***
11b	-1.60	Easy	-1.60	Easy	0.00	Negligible	0.02	0.00		2.10	
11c	-0.96	Easy	-0.93	Easy	-0.02	Negligible	0.02	-1.29		2.72	
11d	0.50	Average	0.63	Difficult	-0.13	Negligible	0.02	-6.87	***	49.39	***
11e	0.94	Difficult	0.94	Difficult	0.00	Negligible	0.02	0.00		2.04	
12a	0.31	Average	0.31	Average	0.00	Negligible	0.02	0.00		0.35	
12b	0.97	Difficult	0.97	Difficult	0.00	Negligible	0.02	0.00		0.03	
13	-0.06	Average	-0.06	Average	0.00	Negligible	0.02	0.00		1.03	
14a	-0.01	Average	-0.11	Average	0.10	Negligible	0.02	5.66	***	38.19	***
14b	0.28	Average	0.18	Average	0.10	Negligible	0.02	5.51	***	33.41	***

Note. ***** = p<.001, ** = p<.01, * = p<.05.**

Table 43. Differential item functioning results by race/ethnicity for the literacy domain

Progression		White Rasch Item Difficulty	Non-white Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
15a	-1.73	Easy	-1.62	Easy	-0.10	Negligible	0.01	-7.22	***	49.40	***
15b	-2.20	Easy	-2.17	Easy	-0.02	Negligible	0.01	-1.66		3.82	
15c	0.94	Difficult	0.94	Difficult	0.00	Negligible	0.01	0.00		0.47	
15d	1.87	Difficult	1.69	Difficult	0.18	Negligible	0.02	9.42	***	86.50	***
16a	-2.13	Easy	-2.05	Easy	-0.08	Negligible	0.01	-6.23	***	32.95	***
16b	-1.01	Easy	-1.01	Easy	0.00	Negligible	0.01	0.00		4.63	
17a	-0.37	Average	-0.24	Average	-0.13	Negligible	0.02	-8.19	***	70.71	***
17b	-1.19	Easy	-1.19	Easy	0.00	Negligible	0.01	0.00		1.47	
18a	0.78	Difficult	0.81	Difficult	-0.03	Negligible	0.01	-1.94		6.55	
18b	-2.18	Easy	-2.18	Easy	0.00	Negligible	0.01	0.00		1.98	
18c	1.07	Difficult	1.07	Difficult	0.00	Negligible	0.01	0.00		0.05	
18d	3.14	Difficult	2.75	Difficult	0.39	Negligible	0.03	14.62	***	100.00	***
18e	3.01	Difficult	2.73	Difficult	0.28	Negligible	0.03	10.65	***	100.00	***
19a	-2.88	Easy	-2.88	Easy	0.00	Negligible	0.01	0.00		8.62	
19b	0.10	Average	0.10	Average	0.00	Negligible	0.01	0.00		4.77	
19c	3.12	Difficult	2.72	Difficult	0.41	Negligible	0.03	15.65	***	100.00	***

Note. ***** = $p<.001$, ** = $p<.01$, * = $p<.05$.**

Table 44. Differential item functioning results by race/ethnicity for the mathematics domain

Progression		White Rasch Item Difficulty	Non-white Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint se	Rasch-Welch t	p	Mantel-Haenszel x^2	p
20a	-0.87	Easy	-0.87	Easy	0.00	Negligible	0.01	0.00		5.53	
20b	-0.92	Easy	-0.92	Easy	0.00	Negligible	0.01	0.00	**	1.86	***
20c	-0.98	Easy	-0.98	Easy	0.00	Negligible	0.01	0.00		0.01	
20d	2.35	Difficult	2.14	Difficult	0.21	Negligible	0.03	7.56		41.37	
20e	2.47	Difficult	2.22	Difficult	0.25	Negligible	0.03	9.43	**	82.58	**
20f	2.09	Difficult	1.87	Difficult	0.22	Negligible	0.03	8.62	*	51.89	
21a	-4.14	Easy	-4.07	Easy	-0.07	Negligible	0.01	-4.81	***	30.87	***
21b	-0.79	Easy	-0.70	Easy	-0.09	Negligible	0.01	-6.09	***	48.23	***
22a	-0.16	Average	-0.11	Average	-0.05	Negligible	0.01	-3.83	***	14.39	***
22b	0.82	Difficult	0.76	Difficult	0.05	Negligible	0.02	3.38	***	3.36	***
22c	1.08	Difficult	0.99	Difficult	0.09	Negligible	0.02	5.02		20.88	
23	-0.74	Easy	-0.67	Easy	-0.07	Negligible	0.01	-5.20		26.31	

Note. ***** = p<.001, ** = p<.01, * = p<.05.**

Table 45. Demographic characteristics of the sample for the DIF analyses

Characteristic	Category	Percent
Classroom Type	3 Year Old Preschool	34.5
	4 Year Old Prekindergarten	65.5
Gender	Female	46.8
	Male	53.2
Primary Language Spoken in the Home	English	87.9
	Spanish	10.2
	Other	1.9
Dual-Language Learner	Yes	13.8
	No	86.2
Race/Ethnicity	White	42.3
	Black	23.1
	Asian	2.2
	Native American/Pacific Islander	2.1
	Hispanic	27.1
	Two or More Races	3.2
Qualifies for National School Lunch Program	Yes	24.0
	No	76.0
Disability Status	Has an IEP	21.1
	No IEP	78.9
	Previously had an ISFP	1.9

Table 46. Differential item functioning results by IEP Status for the social–emotional domain

Progression		No IEP Rasch Item Difficulty	IEP Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
1a	0.05	Average	0.00	Average	0.05	Negligible	0.01	4.97	***	18.48	***
1b	-0.22	Average	-0.27	Average	0.05	Negligible	0.01	4.96	***	17.16	***
1c	-0.67	Easy	-0.69	Easy	0.03	Negligible	0.01	3.32	***	4.71	*
2a	-2.00	Easy	-2.42	Easy	0.42	Negligible	0.01	48.67	***	100.00	***
2b	0.38	Average	0.42	Difficult	-0.04	Negligible	0.01	-3.83	***	1.11	
2c	0.59	Difficult	0.75	Difficult	-0.16	Negligible	0.01	-16.70	***	37.29	***
2d	0.67	Difficult	0.78	Difficult	-0.11	Negligible	0.01	-12.00	***	100.00	***
3a	0.42	Average	0.58	Difficult	-0.16	Negligible	0.01	-16.90	***	53.92	***
3b	0.74	Difficult	0.96	Difficult	-0.22	Negligible	0.01	-23.90	***	100.00	***

Note. ***** = p<.001, ** = p<.01, * = p<.05.**

Table 47. Differential item functioning results by IEP Status for the physical domain

Progression		No IEP Rasch Item Difficulty	IEP Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint se	Rasch-Welch t	p	Mantel-Haenszel x^2	p
4	-0.82	Easy	-1.01	Easy	0.18	Negligible	0.18	18.67	***	100.00	***
5	0.23	Average	0.14	Average	0.09	Negligible	0.09	9.28	***	100.00	***
6	0.59	Difficult	0.49	Average	0.10	Negligible	0.10	10.31	***	100.00	***
7a	-0.52	Easy	-0.37	Average	-0.15	Negligible	-0.15	-14.90	***	100.00	***
7b	0.53	Difficult	0.76	Difficult	-0.23	Negligible	-0.23	-24.00	***	100.00	***

Note. ***** = p<.001, ** = p<.01, * = p<.05.**

Table 48. Differential item functioning results by IEP Status for the language domain

Progression		No IEP Rasch Item Difficulty	IEP Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint se	Rasch-Welch t	*p*	Mantel-Haenszel x^2	*p*
8a	-0.43	Average	-0.91	Easy	0.48	Intermediate	0.01	45.66	***	100.00	***
8b	-1.33	Easy	0.15	Average	0.37	Negligible	0.01	35.29	***	100.00	***
9a	0.48	Average	0.11	Average	0.28	Negligible	0.01	26.08	***	100.00	***
9b	-0.12	Average	-0.08	Average	-0.18	Negligible	0.01	-17.60	***	100.00	***
9c	0.78	Difficult	-0.16	Average	-0.22	Negligible	0.01	-34.30	***	100.00	***
9d	0.63	Difficult	-0.21	Average	-0.44	Negligible	0.01	-45.00	***	100.00	***
10a	0.25	Average	-0.04	Average	-0.09	Negligible	0.01	-9.17	***	100.00	***
10b	-0.19	Average	0.03	Average	0.07	Negligible	0.01	6.43	***	100.00	***

Note. ****** = p<.001, ** = p<.01, * = p<.05.***

Table 49. Differential item functioning results by IEP Status for the cognitive domain

Progression		No IEP Rasch Item Difficulty	IEP Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
11a	-0.48	Average	-0.48	Average	0.00	Negligible	0.01	0.00		1.26	
11b	-0.43	Average	-0.54	Easy	0.11	Negligible	0.01	9.75	***	57.35	***
11c	-0.36	Average	-0.31	Average	-0.05	Negligible	0.01	-4.69	***	11.73	***
11d	0.10	Average	-0.15	Average	0.25	Negligible	0.01	22.67	***	100.00	***
11e	0.37	Average	0.42	Average	-0.04	Negligible	0.01	-3.92	***	0.53	
12a	0.24	Average	0.39	Average	-0.15	Negligible	0.01	-13.80	***	100.00	***
12b	-0.03	Average	0.12	Average	-0.15	Negligible	0.01	-14.40	***	100.00	***
13	0.24	Average	0.04	Average	0.20	Negligible	0.01	18.34	***	100.00	***
14a	-0.09	Average	0.10	Average	-0.19	Negligible	0.01	-15.80	***	100.00	***
14b	0.45	Average	0.45	Average	0.00	Negligible	0.01	0.00		0.20	

Note. ***** = p<.001, ** = p<.01, * = p<.05.**

Table 50. Differential item functioning results by IEP Status for the literacy domain

Progression		No IEP Rasch Item Difficulty	IEP Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
15a	-0.75	Easy	-0.60	Easy	-0.15	Negligible	0.01	-17.20	***	100.00	***
15b	-0.70	Easy	-0.50	Average	-0.20	Negligible	0.01	-22.10	***	100.00	***
15c	0.54	Difficult	0.67	Difficult	-0.13	Negligible	0.01	-13.30	***	1.40	
15d	1.03	Difficult	0.83	Difficult	0.20	Negligible	0.01	13.74	***	100.00	***
16a	-0.82	Easy	-1.19	Easy	0.37	Negligible	0.01	53.23	***	100.00	***
16b	0.12	Average	0.02	Average	0.10	Negligible	0.01	11.08	***	100.00	***
17a	-0.10	Average	-0.36	Average	0.25	Negligible	0.01	27.83	***	100.00	***
17b	-0.01	Average	0.05	Average	-0.06	Negligible	0.01	-6.24	***	2.88	
18a	0.18	Average	0.41	Average	-0.23	Negligible	0.01	-25.80	***	100.00	***
18b	-0.79	Easy	-0.77	Easy	-0.02	Negligible	0.01	-2.48	**	29.02	***
18c	0.18	Average	0.46	Average	-0.28	Negligible	0.01	-30.40	***	100.00	***
18d	1.31	Difficult	0.97	Difficult	0.34	Negligible	0.02	18.03	***	100.00	***
18e	1.50	Difficult	1.08	Difficult	0.42	Negligible	0.02	17.09	***	100.00	***
19a	-1.54	Easy	-1.54	Easy	0.00	Negligible	0.01	0.00		100.00	***
19b	-0.96	Easy	-0.90	Easy	-0.06	Negligible	0.01	-9.40	***	100.00	***
19c	1.09	Difficult	0.78	Difficult	0.31	Negligible	0.02	14.14	***	100.00	***

Note. ***** = p<.001, ** = p<.01, * = p<.05.**

Table 51. Differential item functioning results by IEP Status for the mathematics domain

Progression		No IEP Rasch Item Difficulty	IEP Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel-Haenszel x^2	*p*
20a	-0.33	Average	-0.28	Average	-0.05	Negligible	0.01	-5.96	***	77.84	***
20b	-0.46	Average	-0.34	Average	-0.12	Negligible	0.01	-12.80	***	100.00	***
20c	-0.28	Average	-0.46	Average	0.18	Negligible	0.01	22.58	***	100.00	***
20d	1.18	Difficult	0.82	Difficult	0.35	Negligible	0.03	13.30	***	91.45	***
20e	1.52	Difficult	1.35	Difficult	0.17	Negligible	0.03	6.14	***	21.80	**
20f	1.54	Difficult	1.25	Difficult	0.30	Negligible	0.03	10.76	***	73.38	***
21a	-2.02	Easy	-1.84	Easy	-0.18	Negligible	0.01	-20.70	***	100.00	***
21b	-0.44	Average	-0.75	Easy	0.31	Negligible	0.01	33.71	***	100.00	***
22a	-0.23	Average	-0.11	Average	-0.11	Negligible	0.01	-12.00	***	100.00	***
22b	0.23	Average	0.25	Average	-0.02	Negligible	0.01	-2.33	*	2.13	
22c	0.53	Difficult	0.83	Difficult	-0.30	Negligible	0.01	-24.90	***	99.51	***
23	-1.07	Easy	-1.12	Easy	0.04	Negligible	0.01	5.13	***	1.31	

Note. ***** = p<.001, ** = p<.01, * = p<.05.**

Table 52. Fit statistics for each CFA model

	CFI	TFI	RMSEA	90% LL	90% UL	SRMR
Measurement Models						
Social–Emotional	.986	.980	.095	.093	.097	.009
Physical	.999	.998	.038	.033	.043	.002
Language	.984	.978	.110	.107	.112	.008
Cognitive	.985	.980	.095	.093	.097	.007
Literacy	.999	.999	.078	.077	.079	.054
Mathematics	.999	.999	.029	.027	.030	.029
Confirmatory Factor Analyses						
Model 1	.997	.997	.051	.051	.051	.043
Model 2	.997	.997	.055	.054	.055	.048

Note. CFI - Comparative Fit Index, TLI - Tucker-Lewis Index, 90% LL - Lower Limit, 90% UL - Upper Limit, RMSEA - Root Mean Square Error of Approximation, SRMR - Standardized Root Mean Square Residual.

Table 53. Fit statistics for the measurement invariance analyses by IEP status

	CFI	TFI	RMSEA	90% LL	90% UL	SRMR
Configural Invariance						
No IEP	.988	.987	.067	.067	.067	.063
IEP	.981	.980	.079	.079	.080	.076
Weak Invariance	.985	.984	.074	.074	.074	.071
Strong Invariance	.984	.985	.073	.073	.074	.070
Strict Invariance	.984	.985	.073	.073	.073	.070

Note. CFI - Comparative Fit Index, TLI - Tucker-Lewis Index, 90% LL - Lower Limit, 90% UL - Upper Limit, RMSEA - Root Mean Square Error of Approximation, SRMR - Standardized Root Mean Square Residual.

Table 54. Fit statistics for the measurement invariance analyses by gender

	CFI	TFI	RMSEA	90% LL	90% UL	SRMR
Configural Invariance						
Male	.997	.997	.053	.053	.053	.045
Female	.998	.997	.048	.048	.048	.042
Weak Invariance	.997	.997	.051	.051	.051	.045
Strong Invariance	.997	.997	.050	.049	.050	.044
Strict Invariance	.997	.997	.049	.049	.049	.044

Note. CFI - Comparative Fit Index, TLI - Tucker-Lewis Index, 90% LL - Lower Limit, 90% UL - Upper Limit, RMSEA - Root Mean Square Error of Approximation, SRMR - Standardized Root Mean Square Residual.

Table 55. Fit statistics for the measurement invariance analyses by primary language

	CFI	TFI	RMSEA	90% LL	90% UL	SRMR
Configural Invariance						
English	.997	.997	.051	.051	.051	.043
Spanish	.992	.992	.062	.061	.063	.060
Weak Invariance	.993	.993	.077	.077	.077	.056
Strong Invariance	.996	.996	.057	.057	.057	.049
Strict Invariance	.996	.996	.057	.057	.057	.050

Note. CFI - Comparative Fit Index, TLI - Tucker-Lewis Index, 90% LL - Lower Limit, 90% UL - Upper Limit, RMSEA - Root Mean Square Error of Approximation, SRMR - Standardized Root Mean Square Residual.

Table 56. Fit statistics for the measurement invariance analyses by race/ethnicity

	CFI	TFI	RMSEA	90% LL	90% UL	SRMR
Configural Invariance						
White	.998	.998	.045	.045	.045	.038
Children of Color	.995	.995	.054	.054	.055	.049
Weak Invariance	.994	.994	.074	.073	.074	.069
Strong Invariance	.997	.997	.053	.053	.054	.050
Strict Invariance	.997	.997	.053	.053	.054	.051

Note. CFI - Comparative Fit Index, TLI - Tucker-Lewis Index, 90% LL - Lower Limit, 90% UL - Upper Limit, RMSEA - Root Mean Square Error of Approximation, SRMR - Standardized Root Mean Square Residual.

Glossary

Color Band – Under each developmental progression, color bands provide the range of widely held expectations for a particular group (age, class, or grade) of children.

Developmental Progression – Each developmental progression provides a continuum of behavioral expectations under a given objective for birth through third grade. Teachers rate children's abilities according to the scale provided and the evidence gathered for each child.

Dimension – Dimensions are the narrowest levels of *GOLD*®'s hierarchy. Objectives are made up of one or more dimensions that further define the associated observable behaviors.

Domain – Domains are broadest levels of *GOLD*®'s hierarchy. They are the comprehensive areas of development and learning that teachers assess using *GOLD*®. The ten domains include four developmental domains (social–emotional, physical, language, and cognitive), five content domains (literacy, mathematics, science and technology, social studies, and the arts), and one additional domain for English- and dual-language learners (English language acquisition).

Objective – Each domain is made up of a set of objectives designed to guide teachers through the assessment process. Objectives enable teachers to link observable behavior to essential developmental and early learning requirements. *GOLD*® has 38 objectives in total, which Teaching Strategies calls the "objectives for development and learning."

Raw Score – A raw score represents a child's placement, or current status level, on a given developmental progression. Summing the raw scores within a domain will provide a child's total raw score for that domain.

Scaled Score – Scaled scores convert total raw scores to a common scale. These scores can be used to track children's growth and development over time. The *GOLD*® uniform scale ranges from 0 to 1000 for six areas of development and learning: social-emotional, physical, language, cognitive, literacy, and mathematics. Unlike with raw scores, with scaled scores one point of growth represents the "same amount" of development within a given domain.

Widely Held Expectations (WHE) – WHE are the range of knowledge, skills, and abilities that children of a particular age, class, or grade typically demonstrate over a given year of life or from the beginning to the end of a program year.